EFFICIENT NONPARAMETRIC AND SEMIPARAMETRIC REGRESSION METHODS WITH APPLICATION IN CASE-CONTROL STUDIES

A Dissertation

by

SHAHINA RAHMAN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Raymond J. Carroll
Co-Chair of Committee,	Yanyuan Ma
Committee Members,	Ursula Müller-Harknett
	Roger Smith
	Bani K. Mallick
Head of Department,	Valen E. Johnson

August 2015

Major Subject: Statistics

ABSTRACT

Regression Analysis is one of the most important tools of statistics which is widely used in other scientific fields for projection and modeling of association between two variables. Nowadays with modern computing techniques and super high performance devices, regression analysis on multiple dimensions has become an important issue. Our task is to address the issue of modeling with no assumption on the mean and the variance structure and further with no assumption on the error distribution. In other words, we focus on developing robust semiparametric and nonparamteric regression problems. In modern genetic epidemiological association studies, it is often important to investigate the relationships among the potential covariates related to disease in case-control data, a study known as "Secondary Analysis". First we focus to model the association between the potential covariates in univariate dimension nonparametrically. Then we focus to model the association in mulivariate set up by assuming a convenient and popular multivariate semiparametric model, known as Single-Index Model. The secondary analysis of case-control studies is particularly challenging due to multiple reasons (a) the case-control sample is not a random sample, (b) the logistic intercept is practically not identifiable and (c) misspecification of error distribution leads to inconsistent results. For rare disease, controls (individual free of disease) are typically used for valid estimation. However, numerous publication are done to utilize the entire case-control sample (including the diseased individual) to increase the efficiency. Previous work in this context has either specified a fully parametric distribution for regression errors or specified a homoscedastic distribution for the regression errors or have assumed parametric forms on the regression mean.

In the first chapter we focus on to predict an univariate covariate Y by another potential univariate covariate $\boldsymbol{X}$ neither by any parametric form on the mean function nor by any distributional assumption on error, hence addressing potential heteroscedasticity, a problem which has not been studied before. We develop a tilted Kernel based estimator which is a first attempt to model the mean function non-parametrically in secondary analysis. In the following chapters, we focus on i.i.d samples to model both the mean and variance function for predicting Y by multiple covariates $\boldsymbol{X}$ without assuming any form on the regression mean. In particular we model Y by a single-index model $m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a single-index vector and m is unspecified. We also model the variance function by another flexible single index model. We develop a practical and readily applicable Bayesian methodology based on penalized spline and Markov Chain Monte Carlo (MCMC) both in i.i.d set up and in case-control set up. For efficient estimation, we model the error distribution by a Dirichlet process mixture models of Normals (DPMM). In numerical examples, we illustrate the finite sample performance of the posterior estimates for both i.i.d and for case-control set up. For single-index set up, in i.i.d case only one existing work based on local linear kernel method addresses modeling of the variance function. We found that our method based on DPMM vastly outperforms the other existing method in terms of mean square efficiency and computation stability. We develop the single-index modeling in secondary analysis to introduce flexible mean and variance function modeling in case-control studies, a problem which has not been studies before. We showed that our method is almost 2 times efficient than using only controls, which is typically used for many cases. We use the real data example from NIH-AARP study on breast cancer, from Colon Cancer Study on red meat consumption and from National Morbidity Air Pollution Study to illustrate the computational efficiency and stability of our methods.

DEDICATION

to Sourav for the entire journey of PhD,

and to

Ma and Abbu for believing in me.

ACKNOWLEDGEMENTS

This work wouldn't have been possible without the enormous guidance and encouragement of my PhD Advisor, Raymond J. Carroll. I am immensely grateful to my co-Advisor, Yanyuan Ma for her huge inspiration and support in my research. I hereby acknowledge that the second chapter of my Thesis is already published in Springer in the reputed journal of "Statistics in Bioscience" in 2014 under the name of "A Tilted Kernel Estimator for Nonparametric Regression in the Secondary Analysis of Case-Control Studies". I would also like to thank the IT department head, Henrik Schmiediche and his team for giving me proper computational instructions and thorough technical supports. I must thank our Department Head, Val Johnson for providing us extraordinary financial and constant professional support. I would also like to thank Suhasini Subba Rao, Urshi Mueller, Jeffrey Hart and Mohsen Pourahmadi for several important comments and tremendous motivations which helped me to complete my research.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

1. INTRODUCTION

1.1 Nonparametric Efficient Regression for Secondary Analysis in Case-Control Studies

Case-control designs consist of a large fraction of cases (i.e diseased individuals) and a comparable sample of controls (i.e disease-free individuals). Because they over-sample the cases, the resulting sample is not representative of the entire population. It is an efficient and regularly used study design for rare diseases,such as most cancers, and is widely used in genome-wide association studies (GWAS). The *primary* analysis of case-control association studies is to model the risk of disease D by covariates, denoted here by Y and X. *Secondary* analysis of case-control studies models the association between the covariates Y and X for the general population. To the best of our knowledge, there are no methods for nonparametric regression estimation of a mean function in the literature in this context. The goal of this paper is to regress Y on X without any parametric form on the regression function, without assuming homoscedasticity and with no distributional assumptions on Y. Four types of major analysis have been conducted to assess the effects of Y given X using data from case-control studies: (a) use only the controls; (b) use only the cases; (c) use all the data and without taking into account the case-control sampling design; and (d) use all the data but take the sampling design into account. For rare disease, i.e., the disease rate less than 1% in the population, controls can be regarded as almost a random sample from the population, and hence it is common practice to regress Y on X among the controls only. Analysis based on the cases only or using the entire case-control sample without taking the sampling design into account leads to serious bias because of the over-sampling of the cases.

In recent years, the secondary analysis of case-control studies has received increasing attention, with the intention of using all the data while taking the sampling design into account, and thus improving efficiency of analysis compared to using the controls only. See, for example, Jiang, et al. (2006), Lin and Zeng (2006, 2010), Chen, et al. (2008), Monses, et al. (2009), and Wei, et al. (2013). These papers all have parametric models for the regression function, some model the distribution of Y given X in a parametric manner, and other insist that the regression is homoscedastic. Existing parametric models are not robust to either misspecification of the distribution of Y given X (Wei, et al., 2013), heteroscedasticity or to misspecification of the parametric regression model.

Our focus here is on nonparametric regression, a problem that has not been discussed in the case-control literature. Since ordinary kernel regression based on all the subjects is highly biased, we develop an novel tilted or adjusted kernel-type estimator that allows us to use all the data in order to increase efficiency without introducing bias. While doing so we relax the assumption of any distributional form for the regression model and do not assume homoscedasticity, things assumed by current methods. If the disease rate *in the population* is known, we show that our tilted kernel-type estimator is consistent and is asymptotically normally distributed.

Importantly, the disease rate in the population is typically unknown. Without any assumptions about the distribution of (Y, X) *in the population*, the population disease rate cannot be estimated (Prentice and Pyke, 1979; Chatterjee and Carroll, 2005), and thus $E(Y|X)$ *is not estimable nonparametrically from the case-control study.* In this context, many researchers make a rare disease approximation (see references in Section 3.1.3), in which the regression among the controls is approximately consistent for the regression function. We show how to modify our tilted kernel estimator to be consistent for the regression among the controls, and also more efficient than using

only the controls.

Primarily, we have considered the case of nonparametrically estimating $E(Y|X)$ when no assumptions about the distribution of Y given X are made, including homoscedasticity. Section 3.2.2 describes methodology in the rare case that the disease rate in the population, π_1, is known. Sections B.1 - 3.2.2 describe methodology in the far more common case that π_1 is unknown. In this common case, our simulations show conclusively that our tilted kernel estimator defined in Section 3.2.2 is the more efficient.

In Section 2.4, we considered the case that the disease rate in the population is unknown, and when one is willing to specify a distribution for Y given X up to a function $\mu(X)$ and other parameters, using a local likelihood method along with profiling methods. We displayed the method for when Y is binary with mean $H\{\mu(X)\}$. However, we emphasized two important points: (a) such methods are not consistent if the parametric model is misspecified; and (b) it is likely that the logistic intercept θ_0 will be very difficult to estimate numerically, and a rare disease approximation will improve computational performance (since it eliminates θ_0) while entailing little if any bias.

Ours is the first paper to consider nonparametric regression in the secondary analysis of case-control studies. We have focused on the case of scalar X, and discovered a tilted kernel approach for estimation. With this tilted kernel function, extensions to multivariate X are surely possible, including multivariate kernel regression (Ruppert and Wand, 1994), additive models, etc. However, with multivariate X or higher dimensional covariates, multivariate kernel regression suffers from "Curse of Dimensionality". So next we focus on a popular and relatively simpler multivariate regression tool known as "Single-Index Model" to model the regression function. To make the model more flexible, we also model the variance function by another

single-index model for addressing potential heteroscedasticity. In i.i.d framework, estimating the mean function by a single-index model is well established, if we ignore heteroscedasticity. Lian et al. (2015) proposed a two-stage semiparametric Kernel approach to model the mean and variance function separately. We take a Bayesian approach based on finite mixture of normals which could provide a stable and practical estimates of mean and variance function both on i.i.d set up and then on case-control set up.

1.2 Single-Index Model for Mean and Variance Function for i.i.d samples

The single-index model is an important tool in multivariate nonparametric regression with useful and extensive application in various fields like econometric and biometrics. It reduces the dimensionality from multivariate covariates to an univariate predictor $\boldsymbol{X}^{\mathbf{T}}\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is a dimension reduction index. Hence, a single index model avoids the curse of dimensionality (Bellman, 1961) while still capturing the important features in high-dimensional data. The single index model essentially generalizes linear regression by replacing the linear combination with an unknown univariate link function $m(\boldsymbol{X}^{\mathbf{T}}\boldsymbol{\theta})$. So the model can retain the flexibility of nonparametric regression model with dimension reduction ability. In this paper we allow the regression model to be heteroscedastic, so that the variance function depends on another single-index $\boldsymbol{X}^{\mathbf{T}}\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ is a dimension reduction index for the variance function.

Various methods are already well established to model the mean function, if we ignore the potential heteroscedasticity. Ichimura (1993) and Härdle et al. (1997) used kernel smoothing. Carroll et al. (1997) used local linear methods. Stoker (1986) and Härdle and Stoker (1989) used the average derivative method. Even with sophisticated bandwidth selection or iterative improved techniques, the numerical

instability of the above approaches persists and the potential weaknesses are discussed in details by Yu and Ruppert (2002). Yu and Ruppert (2002) proposed a penalized spline estimation procedure. Xia and Härdle (2006) integrated the dimension reduction technique with minimum average variance estimation (MAVE, Xia, 2002). Ma and Zhu (2012) developed a semiparametic dimension reduction method in a multiple-index structure and studied the single-index model as a special case. A semi-Bayesian model based on P-spline and a random walk Metropolis algorithm was proposed by Antoniadis et al. (2004) only to model the mean function who remarked that "A Bayesian approach offers a relatively easy to implement method with a hope of more stable estimates, especially for small or moderate sample size". However, they used generalized cross validation (GCV) to choose the smoothness paramter for the B-splines.

None of the above works in single-index modeling considered estimation of the variance function, which can play a crucial role to construct the confidence intervals for the mean function (Cai and Wang, 2008). Several works in last three decades have shown that the estimation of variance function has improved the model fit. Box and Hill (1974) improved the model fit in the study of kinetic rate parameters using variance function. In offline quality control, Box and Meyer (1986) emphasizes not only on the mean response but also in its variability to improve understanding the model. Box, 1986; Box and Ramirez, 1986 advocated to employ effective variance function estimation to account for the heteroscedasticity. Teschendorff and Windschwenter (2012) recognized in cancer genomics that variability can be a predictor of disease phenotypes. A number of publication has been done based on parametric approaches to model the variance function (Bickel, 1978; Carroll, 1982; Carroll and Ruppert, 1982; Davidian and Carroll, 1987). Caroll and Härdle (1989), Fuller and Rao (1978) and Hall and Carroll (1989) modeled the variance function nonparametri-

cally. Ma et al. (2006) studied semiparametric efficiency in heteroscedastic partially linear models where X is scalar. Van Keilegom and Wang (2010) studied a general class of location-dispersion regression models, including semiparametric quantile heteroscedastic regression. Ma and Zhu (2013) developed a doubly robust and efficienct estimators of the mean parameters. For the single-index model, recently Lian et al. (2014) introduced a semiparametric Kernel based efficient estimator for estimating the mean and variance function simultaneously.

The primary goal of this work is to introduce a fully Bayesian approach for a heteroscedastic nonparamteric single-index model. A fully Bayesian approach in this problem can have the potential to achieve large gains in efficiency of estimation compared to the existing Kernel method developed by Lian et al. (2014). In this article, the density of the scaled error is modeled by two approaches, (a) the normal distribution, and (b)a Dirichlet process mixture of normals (DPMN). Implementing the MCMC algorithm under normal distribution is straightforward and provides consistent estimation of the parameters. However, to ensure a more flexible representation of the scaled errors we model the error by DPMN, which potentially increases the efficency in estimation under heavy tail distribution or in the presence of potential outliers. Modeling the density of interest by a flexible location-scale mixture of normal induced by a Dirichlet process provides an efficient alternative to nonparametric modeling in Bayesian set up (Ferguson, 1973, Escobar and West, 1995). Griffin and Steel (2010), Pelenis (2014), Sarkar et al. (2014) used the flexible model based on Dirichlet process to achieve more efficiency in estimating the model to gain considerable mean squared efficiency.

1.3 Single-Index Modeling in Secondary Analysis for Case-Control Studies

Primarily case-control samples are used to study the relationship between rare diseases (like most cancers) and the covariates. It is a popular and efficient design to understand the risk factors for cancers and other rare diseases. In population there are two groups, one with the disease, called *cases*, and another without the disease, called *controls*. Random samples of comparable sizes are separately drawn from each group to form the case-control samples. Data on various covariates are then collected in the restrospective fashion. Due over representation of the cases in the sample, case-control sample is not a representative of the entire population. Therefore, generally, the case-control data set cannot be used as if it were a random sample from the true population. In this study we model the association between the potential covariates which are related with the disease. Indeed, unless disease is independent of Y given $\boldsymbol{X}$, the regression of Y on $\boldsymbol{X}$ based on the case-control sample *as it is*, will lead to a relationship different from that in the true population. The goal of this paper is to model a covariate Y by a set of multiple covariates $\boldsymbol{X}$ nonparametrically including all the case-control sample, a problem which has not been addressed previously in the case-control set up.

The standard method of the primary analysis of case-control data involves logistic regression modeling of the disease outcome as a function of the covariates of interest. Epidemiologic researchers popularly use controls from case-control studies to examine the interrelationship between certain covariates themselves, known as *secondary analysis*. Such studies has received increasing attention, where it is often of interest to investigate the effect of Age or Alcohol intake or Fat density, not only on the primary disease outcome, but also on other secondary factors like BMI (Body Mass Index).

This study is particularly challenging due to multiple reasons, (a) case-control sample is biased due to disproportionate number of cases, (b) logistic intercept of the prospective disease model is not practically identifiable, (c) mean function is modeled with no parametric assumption and (d) misspecification of error distribution leads to inconsistent results. In this paper we address all of the above challenges by modeling Y by a generalized linear model $m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})$, known as single-index model where m is unknown and $\boldsymbol{\theta}$ represents the index or the directional vectors.

Single-index model is an efficient and popular nonparametric multivariate method extensively used in econometrics and biomedical fields. For i.i.d case, a number of publication has been done to model the index vector efficiently. For example, Lian, et al. (2015) developed a semiparametric efficient estimation Kernel method considering single-index model for both mean and standard deviation function. Modeling both mean and standard deviation function by single-index model, Rahman, et al. (unpublished) developed an efficient Bayesian method based of finite mixture of Normal based on robust MCMC which doesn't depend on the initial values. Drawing inspiration from our work in second chapter, we seek to develop a semiparametric efficient and robust method to model Y on multivariate $\boldsymbol{X}$ for the case-control set up. We also attempt to model the variance function by another single-index model and the error distribution is modeled by finite mixture of normals. This particular study is not addressed before in secondary analysis.

In recent years, the secondary analysis of case-control studies has received increasing attention, with the intention of using all the data while taking the sampling design into account, and thus improving efficiency of analysis compared to using the controls only. See, for example, Jiang, et al (2006), Lin and Zeng (2006, 2010), Chen, et al. (2008), Monses, et al. (2009), Wei, et al. (2013), Ma and Carroll (2015). These papers all have parametric models for the regression function, some model the dis-

tribution of Y given $\boldsymbol{X}$ in a parametric manner, and other insist that the regression is homoscedastic. Rahman (2014) introduced a Kernel based efficient robust method including all the case-control data for modeling univariate Y given univariate X. However, the method of Rahman (2014) suffers from "curse of dimensionality" if we considered to regress Y on multivariate $\boldsymbol{X}$.

In our last chapter, we present our work on modeling a multivariate nonparametric regression problem in case-control studies which has not been addressed before. We attempt to model a nonparametric multivariate regression problem by Single-index model where we regress Y on a lower dimension$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}$ nonparametrically. This work is also unique because we model the variance by another single-index model based on $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma}$ which has been only previously addressed by Lian et al. (2015) and Rahman et al.(2015) in i.i.d set up. To ensure for flexible modeling, we model the distribution of the error by a finite mixture of normals. The method of Lian et al.(2015) based on semiparametric kernel regression is efficient but the estimation depends heavily on the initial value in estimating equation. In simulation study we showed that when the error is away from normal distribution, the estimates based on normal likelihood lacks efficiency with respect to using only controls. However, the method based on mixture of normal achieves almost 2 times more efficiency than that of using only controls in estimating the mean and the variance function.

2. NONPARAMETRIC REGRESSION IN THE SECONDARY ANALYSIS OF THE CASE-CONTROL STUDIES

2.1 Framework

2.1.1 Background

Let D denote case-control disease status with $D = 1$ denoting a case and $D = 0$ denoting a control. Let Y and X be two univariate continuous covariates for D. For a case-control study with a total of n subjects, the case-control sample consist of $(D_i, Y_i, X_i), i = 1, 2, ..., n$ with n_0 controls and n_1 cases. Let the unknown probability of disease in the population be π_1 and thus $\pi_0 = 1 - \pi_1$ is the non-diseased rate. We assume the underlying nonparametric regression model for Y on X to be

$$Y = \mu(X) + \epsilon, \tag{2.1}$$

where $E(\epsilon|X) = 0$, while the distribution of ϵ given X is unspecified and may be heteroscedastic. As is standard practice, the primary analysis relating Y and X to D is the logistic regression model

$$\text{pr}(D = 1|Y, X) = \exp\{\theta_0 + m(Y, X, \theta_1)\}/[1 + \exp\{\theta_0 + m(Y, X, \theta_1)\}], \tag{2.2}$$

where $m(\cdot)$ is an arbitrary known function.

Prentice and Pyke (1979) showed that θ_0 is not identifiable in the logistic regres-

sion for the case-control set up unless the disease rate π_1 is known. Typically the disease rate is not known, however. When we run logistic regression in the case-control sample, θ_1 is consistently estimated, while the intercept is known to converge to $\kappa = \theta_0 + \log(n_1/n_0) - \log(\pi_1/\pi_0)$ defined by Chatterjee and Carroll, 2005. Define $\Omega = (\kappa, \theta_1)$.

2.1.2 Rare Disease Approximation and Motivation

If $\text{pr}(D = 1) = \pi_1$ in the population is unknown, neither θ_0 nor the regression of Y on X is identified if one makes no assumptions about the distribution of Y given X in the population. Even if one makes assumptions about the distribution of Y given X in the population, if π_1 is unknown, θ_0, while being technically identified, is very difficult to estimate numerically. For this reason, it is very common to make a rare disease approximation, by which only we can eliminate the estimation of θ_0. Most case-control studies take place for rare diseases, and if the disease is fairly rare, the distribution of (Y, X) in the population is well approximated by that among the controls.

Many authors in this field have adopted this approximation, a very non-exhaustive list of which includes Piegorsch,et al. (1994), Epstein and Satten (2003), Lin and Zeng (2006), Modan, et al. (2001), Zhao, et al. (2003), Chatterjee, et al. (2005), Kwee, et al. (2007), Yang, et al. (2009), Lin and Zeng (2009), Hu, et al. (2010), Li, et al. (2010), Chen, et al. (2008, 2009, 2013) and Wei, et al. (2013). Indeed, in the case that the regression of Y on X is linear, and the regression errors are normally distributed and homoscedastic, the software *SPREG* of Lin and Zeng (2009) requires that either the disease rate in the population is specified, thus effectively specifying θ_0, or the rare disease approximation is made which eliminates θ_0 from consideration. In our nonparametric regression context, unless the disease rate in the population

is known, neither θ_0 nor the regression of Y on X are identified, and a rare disease approximation appears to be the only means to estimate the true regression function approximately.

When the disease is "rare", we have that $\text{pr}(D = 1|Y, X) = \exp\{\theta_0 + m(Y, X, \theta_1)\}/[1 + \exp\{\theta_0 + m(Y, X, \theta_1)\}] \approx \exp\{\theta_0 + m(Y, X, \theta_1)\}$ and

$$\text{pr}(D = 0|Y, X) = H(d = 0|Y, X, \theta_0, \theta_1) = 1/[1 + \exp\{\theta_0 + m(Y, X, \theta_1)\}] \approx (2.3)$$

2.2 Methodology and Estimation

2.2.1 Main Goals

Let $f_{X,cont}(x)$ and $f_{X,case}(x)$ be the density functions of X among the controls and cases, respectively, and let f_X be the density function of X in the population. Let $K(\cdot)$ be a known symmetric density function. Define $K_h(u) = h^{-1}K(u/h)$, where h is a bandwidth. Our goal is to estimate $\mu(x) = E(Y|X = x)$, either consistently or approximately: as described previously, if the population disease rate π_1 is unknown, $\mu(x)$ is not identified. Also define $\mu_{\text{cont}}(x) = E(Y|X, D = 1)$.

2.2.2 When the Population Disease Rate π_1 is Known

In the very uncommon case that the population disease rate π_1 is unknown, it is possible to estimate $\mu(x) = E(Y|X)$. Since π_1 is known,

$$E(Y|X) = \pi_1 E(Y|X, D = 1) + \pi_0 E(Y|X, D = 0), \tag{2.4}$$

so separate regression among cases and controls cane be used to consistently estimate $\mu(x)$. However, as described in Section 3.2.2, in real world mostly π_1 is unknown, which leads to the regression among the controls only, is not as efficient as the method

we will describe in that Section.

If the disease rate is known, then $(\theta_0, \theta_1, \kappa)$ can be well estimated. Define $\mathcal{K}_{pop}(y, x, \Omega, \theta_0) = 1 + \exp\{\kappa + m(y, x, \theta_1)\}/[1 + \exp\{\theta_0 + m(y, x, \theta_1)\}]$. We "tilt" the usual Nadaraya-Watson kernel estimator as follows. Define

$$\Lambda_{pop}(Y_i, x_0, h, \Omega, \theta_0) = \int \mathcal{K}_{pop}(Y_i, v, \Omega, \theta_0) K_h(v - x_0) f_X(v) dv. \tag{2.5}$$

Then, if π_1 is known, in Theorem B.1 in the Appendix A.2, we show that a consistent estimator of $\mu(x_0)$ using both case and control data is

$$\widehat{M}_h(x_0) = \frac{n^{-1} \sum_{i=1}^n Y_i K_h(X_i - x_0)/\Lambda_{pop}(Y_i, x_0, h, \Omega, \theta_0)}{n^{-1} \sum_{i=1}^n K_h(X_i - x_0)/\Lambda_{pop}(Y_i, x_0, h, \Omega, \theta_0)}. \tag{2.6}$$

2.2.3 When π_1 is Unknown: the Control Only Method

It is well-known that since the entire case-control sample is not a random sample from the true population, but instead highly over-represents the cases, the local average kernel estimator (usual Nadaraya Watson estimator) based on the entire sample given by

$$\widehat{m}_{h,\text{naive}}(x_0) \;=\; n^{-1} \sum_{i=1}^n Y_i K_h(X_i - x_0) / n^{-1} \sum_{i=1}^n K_h(X_i - x_0). \tag{2.7}$$

is highly biased except when Y is independent of the disease status D given X. When π_1 is unknown, the only possibility is to invoke a rare disease approximation that $\mu(x) \approx \mu_{\text{cont}}(x)$. In this case, (2.4) suggests estimation using only the controls. Also Nagerkele, et al. (1995), Jiang, et al. (2006) and Lin and Zeng (2006, 2010) advocated that using only controls data leads to consistent estimation when the disease rate is

small. The local average kernel estimator of $\mu(x_0)$ among the controls is

$$\widehat{m}_{h,\text{cont}}(x_0) \;=\; n^{-1}\sum_{i=1}^{n}(1-D_i)Y_iK_h(X_i-x_0)\big/n^{-1}\sum_{i=1}^{n}(1-D_i)K_h(X_i-x_0) \tag{2.8}$$

Since $\widehat{m}_{h,\text{cont}}(x_0)$ is based on a random sample from the population of controls, (2.8) consistently estimates $\mu_{\text{cont}}(x_0) = E(Y|X=x_0, D=0)$, which with a little algebra is given by $\int yf_{XY}(x_0,y)H(d=0|y,x_0)dy\big/\int f_{XY}(x_0,y)H(d=0|y,x_0)dy$. For rare disease when $\pi_1 \approx 0$, by using (2.4), $\mu_{\text{cont}}(x_0) \approx E(Y|X=x_0) = \mu(x_0)$ as $n\to\infty$ and $h\to 0$.

2.2.4 When π_1 is Unknown: a Tilted More Efficient Estimator

The purpose of this section is to develop a kernel estimator that can utilize all the data efficiently and improve upon (2.8). When π_1 is unknown, as it almost always is, (2.6) is an infeasible estimator because neither θ_0 nor $E(Y|X)$ is identifiable from the case-control sample. However, if the disease is rare, we have that $\text{pr}(D=1|Y,X) = \exp\{\theta_0+m(Y,X,\theta_1)\}/[1+\exp\{\theta_0+m(Y,X,\theta_1)\}] \approx \exp\{\theta_0+m(Y,X,\theta_1)\}$ and equation (4.3)

$$\text{pr}(D=0|Y,X) = H(d=0|Y,X,\theta_0,\theta_1) = 1/[1+\exp\{\theta_0+m(Y,X,\theta_1)\}] \;\approx\; 1.$$

Using this, f_X can be approximated by $f_{X,cont}$ and, $\mathcal{K}_{pop}(y,x,\Omega,\theta_0)$ can be approximated by $\mathcal{K}(y,x,\Omega) = 1+\exp\{\kappa+m(y,x,\theta_1)\}$. In other words, we approximate the original "tilt" Λ_{pop} by a "different tilt" $\int K_h(v-x_0)\mathcal{K}(Y_i,v,\Omega)f_{X,cont}(v)dv$, a quantity unbiasedly estimated by

$$\Lambda_n(Y_i,x_0,h,\Omega) = n_0^{-1}\sum_{j=1}^{n}(1-D_j)\mathcal{K}(Y_i,X_j,\Omega)K_h(X_j-x_0). \tag{2.9}$$

Since $\Omega = (\kappa, \theta_1)$ is estimated consistently by $\widehat{\Omega}$ by ordinary logistic regression of D on (Y, X) (Chatterjee and Carroll, 2005), this leads to our estimator

$$\widehat{m}_h(x_0) = \frac{n^{-1}\sum_{i=1} Y_i K_h(X_i - x_0)/\Lambda_n(Y_i, x_0, h, \Omega)}{n^{-1}\sum_{i=1} K_h(X_i - x_0)/\Lambda_n(Y_i, x_0, h, \Omega)}. \tag{2.10}$$

2.2.5 Algorithm When π_1 is Unknown

When π_1 is unknown, based on the analysis in Section 3.2.2, we propose the following algorithm to implement the weighted adjusted local Nadaraya Watson estimator (2.10).

1. Estimate (κ, θ_1) by $(\widehat{\kappa}, \widehat{\theta_1})$ by ordinary logistic regression of D on (Y, X). This can be done legitimately because it is known that ordinary logistic regression in a case-control study consistently estimates (κ, θ_1). (Prentice and Pyke, 1979; Chatterjee and Carroll, 2005). Denote the estimators of (κ, θ_1) by $\widehat{\Omega}$.

2. Choose a suitable symmetric density function $K_h(\cdot)$ such that it is > 0 for all x in the support and is twice differentiable satisfying the common properties of $\int K(z)dz = 1$, $\int zK(z)dz = 0$, and define $\int z^2K(z)dz = c_1$, $\int K^2(z)dz = c_2$, $\int z^2K^2(z)dz = c_3$.

3. *Bandwidth selection*: In R software, the library *(KernSmooth)* has a quick and simple function *dpill* which implements a direct plug-in approach to bandwidth selection, as described by Ruppert, Sheather and Wand (1995). Since for rare disease the true population is almost same as the population of controls, it is reasonable to use this function among the controls only to estimate the bandwidth. Under the conditions of Theorem 1, this has the correct asymptotic rate of $(n_0 + n_1)^{-1/5}$. It is at least possible in theory to improve upon this, see

Theorem 2, but we found this simple device to be eminently satisfactory in our simulations.

4. Calculate the estimator (2.10) using $\widehat{\Omega}$ and the estimated bandwidth.

2.3 Asymptotic Theory

Our main asymptotic results are stated below, and proved in Appendix B.

Theorem 1 *Define the total sample size as $n = n_0 + n_1$, and assume that $n \to \infty$ and $n_1/n \to 1-c$ $(0 < c < 1)$. Assume that the following sets of regularity conditions hold.*

1. *The density functions $f_X(x)$ and $f_{X,\text{cont}}(x)$ have compact support $\mathcal{S}$ and are > 0 on that support.*

2. *The conditional density function $f_{Y|X}(y)$ is a bounded density function with $\int y^2 f_{Y|X}(y)d(y) < \infty$,*

3. *The density functions $f_{Y|X}(\cdot)$, $f_X(\cdot)$ and $f_{X,\text{cont}}(\cdot)$ are twice continuously and boundedly differentiable with respect to x,*

4. *The kernel density $K(\cdot)$ is twice continuously and boundedly differentiable.*

Then, as $h \to 0$ and $nh \to \infty$,

$$\widehat{M}_h(x_0) = \mu(x_0) + o_p(1)$$

Theorem 2 *Under the conditions of Theorem (1), $\widehat{m}_h(x_0) = \mu_{\text{cont}}(x_0) + o_p(1)$. In addition, as $\pi_1 \to 0$, $\mu_{\text{cont}}(x_0) \approx \mu(x_0)$. Also there exist functions $W(x_0, \Omega, \theta_0)$ and $U(x_0, \Omega, \theta_0)$ defined in Appendix A.1.1, such that $\widehat{m}_h(x_0)$ has asymptotic bias*

$$h^2 W(x_0, \Omega, \theta_0) + \mathcal{O}(n_0 h)^{-1/2} + o\{(nh)^{-1/2} + h^2\},$$

and asymptotic variance equal to

$$(n_0 h)^{-1}\pi_0 U(x_0, \Omega, \theta_0) + o(h/nh).$$

Thus, for $b_0 = \left\{\pi_0 U(x_0, \Omega, \theta_0)/4W^2(x_0, \Omega, \theta_0)\right\}^{1/5}$*, the MSE optimal bandwidth for estimating* $\mu_{\text{cont}}(x_0)$ *is* $h_{opt} = b_0 n_0^{-1/5}$*.*

Theorem 3 *Under the conditions of Theorem 1, let* $\int |K(u)|^{2+\delta} du < \infty$ *and* $E|\epsilon|^{2+\delta}$ *for some* $\delta > 0$*. Then as* $nh \to \infty$*,*

$$(nh)^{-1/2}\left\{\widehat{m}_h(x_0) - \mu_{\text{cont}}(x_0) - h^2 W(x_0, \Omega, \theta_0)\right\} \to Normal\left\{0, U(x_0, \Omega, \theta_0\right\}.$$

In addition as $\pi_1 \to 0$*,* $\mu_{\text{cont}}(x_0) \approx \mu(x_0)$*.*

Corollary 1 *Let the conditions of Theorem 3 hold with the bandwidth* $h_{opt} = b_0 n_0^{-1/5}$*. Then the rate of convergence is optimal and as* $nh \to \infty$*,*

$$n^{2/5}\left\{\widehat{m}_h(x_0) - \mu(x_0)\right\} \to Normal\left\{b_0^{5/2} W(x_0, \Omega, \theta_0), U(x_0, \Omega, \theta_0)\right\}.$$

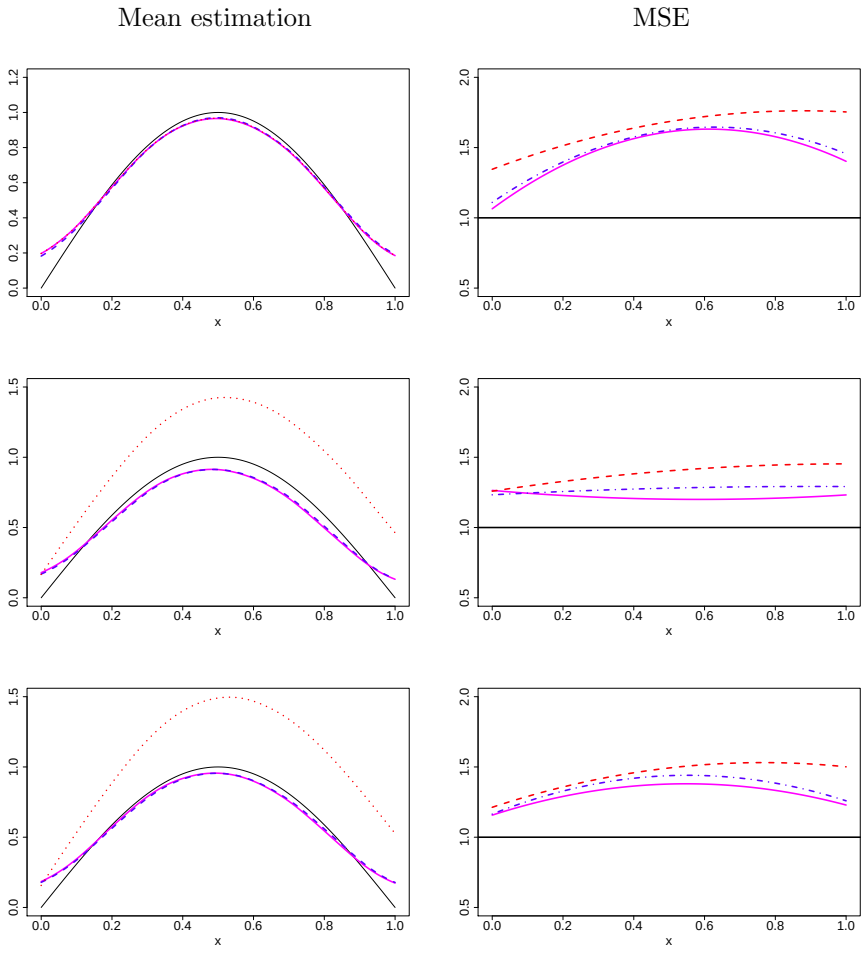

Figure 2.1: The first row is for the case that $\theta_y = 0$ and with a 5% disease rate in the population. The second row is for the case that $\theta_y \neq 0$ and with a 5% disease rate. The second row is for the case that $\theta_y \neq 0$ and with a 1% disease rate. The left panels are the mean estimated functions across all the simulation data sets with $n = 500$: truth (black, solid), using all the data and ignoring the case-control sampling scheme (red, dotted), using only the controls (blue, dashed) and using our new method (magenta, solid). Using only the controls is virtually indistinguishable from our method in terms of the mean. The right panels are the pointwise mean squared error efficiency or our method compared to using only the controls. The dashed red line is for $n = 100$, the dotted blue line is for $n = 300$, and the solid magenta line is for $n = 500$. The solid black line is at 1.0, with values above that indicating that our method is more efficient

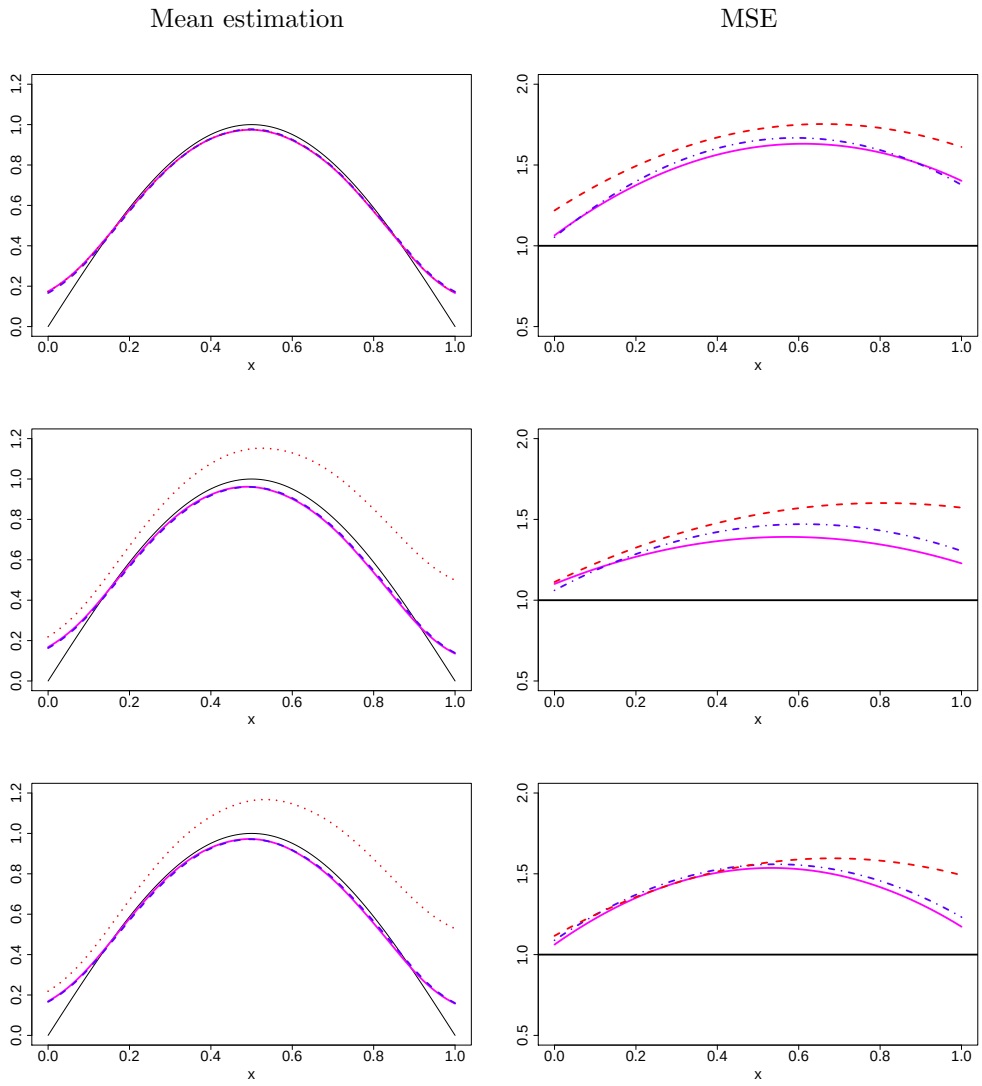

Figure 2.2: The first row is for the case that $\theta_y = 0$ and with a 5% disease rate in the population. The second row is for the case that $\theta_y \neq 0$ and with a 5% disease rate. The second row is for the case that $\theta_y \neq 0$ and with a 1% disease rate. The left panels are the mean estimated functions across all the simulation data sets with $n = 500$: truth (black, solid), using all the data and ignoring the case-control sampling scheme (red, dotted), using only the controls (blue, dashed) and using our new method (magenta, solid). Using only the controls is virtually indistinguishable from our method in terms of the mean. The right panels are the pointwise mean squared error efficiency or our method compared to using only the controls. The dashed red line is for $n = 100$, the dot-dashed blue line is for $n = 300$, and the solid magenta line is for $n = 500$. The solid black line is at 1.0, with values above that indicating that our method is more efficient

2.4 Regression for a Binary Y

To this point, we have made no assumptions about the distribution of Y given X. We continue to assume that π_1 is unknown. As shown by Wei, et al. (2013), even in linear regression, if that distribution is misspecified, methods that make such assumptions are not consistent if the assumptions are misspecified. This is in total contrast to the case of random sampling, where consistency is assured as long as $E(Y|X)$ is specified parametrically and correctly. Insights from random sampling do not pass over to secondary analysis for case control studies.

If one is willing to assume a parametric distribution for Y given X up to a function $\mu(X)$, then local likelihood ideas can be used. However, we caution that the result will not be consistent if the parametric distribution is misspecified.

As a specific example, suppose that Y is binary, and that given X it has mean $H\{\mu(X)\}$, where $H(\cdot)$ is the logistic distribution function. Then, as described by Chen, et al. (2009), Lin and Zeng (2009) and Wei, et al. (2013), the semiparametric efficient retrospective profile likelihood making no assumptions about the distribution of X is given as follows. Define $g\{D, Y, X, \mu(X), \theta_0, \theta_1, \kappa\}$ as

$$g\{d, y, x, \mu(x), \theta_0, \theta_1, \kappa\} = [H\{\mu(x)\}]^y [1 - H\{\mu(x)\}]^{1-y} \frac{\exp[d\{\kappa + m(y, x, \theta_1)\}]}{1 + \exp\{\theta_0 + m(y, x, \theta_1)\}}.$$

Then the semiparametric efficient retrospective profile likelihood making no assumptions about the distribution of X is

$$\mathcal{L}\{D, Y, X, \mu(X), \theta_0, \theta_1, \kappa\} = \frac{g\{D, Y, X, \mu(X), \theta_0, \theta_1, \kappa\}}{\sum_{d=0}^{1} \sum_{t=0}^{1} g\{d, t, X, \mu(X), \theta_0, \theta_1, \kappa\}}.$$

If we want to estimate $\mu(\cdot)$ at x_0, for given $(\theta_0, \theta_1, \kappa)$, we would announce an estimate

$\widehat{M_h}(x_0, \theta_0, \theta_1, \kappa)$ as the solution to

$$\text{argmax}_a \sum_{i=1}^n K_h(X_i - x_0) \log[\mathcal{L}\{D_i, Y_i, X_i, a, \theta_0, \theta_1, \kappa\}].$$

At this point, there are two possibilities.

- Since (κ, θ_1) is estimable from logistic regression of D on (Y, X), get the estimates $(\widehat{\kappa}, \widehat{\theta}_1)$ from that regression. Then define $\widehat{M_h}(x_0)$ as $\widehat{M_h}(x_0, \widehat{\theta}_0, \widehat{\theta}_1, \widehat{\kappa})$, where $\widehat{\theta}_0$ is the solution to the profiled equation

$$\text{argmax}_{\theta_0,} \sum_{i=1}^n K_h(X_i - x_0) \log[\mathcal{L}\{D_i, Y_i, X_i, \widehat{M_h}(X_i, \theta_0, \widehat{\theta}_1, \widehat{\kappa}), \theta_0, \widehat{\theta}_1, \widehat{\kappa}\}].$$

- Alternatively, one could in principle also profile over (κ, θ_1).

At least in principle, it should be the case that $\widehat{M_h}(x_0, \widehat{\theta}_0, \widehat{\theta}_1) \to \mu(x_0)$. However, one needs to be careful with this line of reasoning. Even if the distribution of Y given X is parametrically specified, θ_0 if often very poorly determined, so that $\widehat{\theta}_0$ can take on nearly any value < 0. This is one of the reasons that the software *SPREG* of Lin and Zeng, 2009 in the case that $E(Y|X) = \beta_0 + X\beta_1$ requires that either the disease rate in the population is specified, thus effectively specifying θ_0, or the rare disease approximation is made, and thus eliminating θ_0 from consideration.

2.5 Simulations

2.5.1 Basic Settings

For the typical case when the disease rate in the population is unknown, we conducted simulation studies to evaluate the performance of the standard methods and our new "tilted efficient" method. All simulations are done using the Gaussian kernel density function. We repeated the simulations for sample sizes $n_0 = 100, 300$

and 500, with approximate disease rate of 1% and 5%.

In these simulations, we generated X from the Uniform$(0, 1)$ distribution. We consider a nonlinear regression model $Y = \sin(\pi \mathrm{X}) + \epsilon$ and the prospective logistic model for $\mathrm{pr}(D = 1|Y, X) = H(\theta_0 + \theta_y Y + \theta_x X)$ with $\theta_0 = (-6.00, -5.05)$, $\theta_y = 1.00$, and $\theta_x = 1.80$. By equation (3.3), for $\theta_0 = -6.00$ and $\theta_0 = -5.05$, the disease rate is approximately equals 1% and 5%, respectively. We also consider the case that Y is independent of D given X, by setting the parameters to $\theta_0 = -3.65$, $\theta_y = 0.00$, and $\theta_x = 1$ corresponding to the disease rate $\approx 5\%$, in order to evaluate the methods when Y is independent of the disease status for given X. In each setting, we generated 1000 data sets with an equal number of cases and controls, and repeat the simulation for $n_0 = n_1 = 100, 300, 500$. The bandwidth was estimated as described in Section 2.2.5. We performed two different sets of simulations one with homoscedastic errors and second with heteroscedastic errors. In each set of simulations, we contrasted three methods, (a) ordinary Nadaraya Watson estimator using all the data in (2.7), (b) ordinary Nadaraya Watson estimator using only the controls in (2.8) and (c) our new adjusted Nadaraya Watson estimator in (2.10). The mean square efficiency are smoothed by quadratic regression to illustrate the gain in the large sample efficiency in all the cases.

In the first set of simulations, we generated homoscedastic errors $\epsilon = \mathrm{Normal}(0, \sigma^2)$ with $\sigma^2 = (0.3, 1.00)$. The case $\theta_y = 0.00$ is interesting, since Y is independent of D given X and the three methods gives almost unbiased results, see Figure 2.1. For $\theta_y \neq 0.00$, the estimates obtained by Kernel estimator using all the observations is hugely biased while the other two methods have much less bias. When the disease rate is approximately 1%, both the estimator (2.10) and the control only estimator (2.8) are very close to the true function (Table 2.2). The biases of estimators (2.8) and (2.10) are given in Table 2.1.

Table 2.1: The bias of the new method and the standard method (only controls) in the homoscedastic simulation for $n_0 = 100, 300$ and 500.

		New Method				Controls only			
	x_0	0.19	0.39	0.59	0.79	0.19	0.39	0.59	0.79
	100	-0.024	-0.078	-0.113	-0.092	-0.034	-0.091	-0.115	-0.082
$\sigma = 1.0$	300	-0.029	-0.071	-0.100	-0.089	-0.019	-0.077	-0.098	-0.078
$\pi_1 = 5\%$	500	-0.031	-0.066	-0.100	-0.093	-0.032	-0.064	-0.095	-0.082
	100	-0.003	-0.041	-0.065	-0.047	-0.004	-0.036	-0.065	-0.033
$\sigma = 1.0$	300	-0.006	-0.034	-0.057	-0.043	-0.004	-0.031	-0.055	-0.028
$\pi_1 = 1\%$	500	-0.005	-0.031	-0.052	-0.043	-0.003	-0.041	-0.048	-0.028
	100	-0.005	-0.025	-0.035	-0.031	-0.001	-0.029	-0.033	-0.019
$\sigma = 0.3$	300	-0.008	-0.017	-0.027	-0.025	-0.002	-0.021	-0.025	-0.016
$\pi_1 = 5\%$	500	-0.007	-0.014	-0.025	-0.023	-0.001	-0.017	-0.021	-0.016
	100	-0.003	-0.021	-0.033	-0.025	-0.004	-0.025	-0.029	-0.013
$\sigma = 0.3$	300	-0.006	-0.015	-0.023	-0.020	-0.002	-0.017	-0.021	-0.011
$\pi_1 = 1\%$	500	-0.006	-0.012	-0.019	-0.019	-0.002	-0.014	-0.017	-0.012

We summarize the results for pointwise mean square efficiency in Table 2.2 and Figure 2.1. Our proposed method significantly outperforms the controls only approach in terms of efficiency. When the disease rate is 1%, the overall gain in mean square efficiency becomes is approximately 1.5.

2.5.2 Heteroscedastic Errors

In the second set of simulations, we generated heteroscedastic errors. The same distribution of ϵ was used, except that ϵ was multiplied by $(1 + X^2)^{3/4}/2$ in all the cases, so that the $\text{var}(\epsilon|X) = (1 + X^2)^{3/4}/4$. The results for the heteroscedastic case are summarized in Table 2.3 and 2.4. The results are very similar to the results of the homoscedastic case. Figure 2.2 shows both new method and the controls are almost unbiased for heteroscedastic errors. It also shows that the mean square efficiency of our new method (2.10) is even larger than in the homoscedastic case.

Table 2.2: The mean square efficiency of the new method over the standard method (only controls) in the homoscedastic simulation for $n_0 = 100, 300$ and 500.

		MSE Efficiency					
	X_0	0.01	0.19	0.39	0.59	0.79	0.99
	100	1.137	1.349	1.553	1.403	1.331	1.410
$\sigma = 1.0$	300	1.054	1.358	1.363	1.223	1.143	1.342
$\pi_1 = 5\%$	500	1.026	1.346	1.348	1.117	1.021	1.272
	100	1.187	1.401	1.446	1.498	1.371	1.432
$\sigma = 1.0$	300	1.075	1.490	1.439	1.357	1.254	1.198
$\pi_1 = 1\%$	500	1.042	1.437	1.456	1.274	1.178	1.169
	100	1.019	1.517	1.722	1.563	1.525	1.269
$\sigma = 0.3$	300	0.971	1.435	1.714	1.393	1.269	1.141
$\pi_1 = 5\%$	500	0.971	1.479	1.714	1.309	1.201	1.122
	100	1.009	1.556	1.699	1.492	1.368	1.256
$\sigma = 0.3$	300	0.995	1.527	1.739	1.506	1.349	1.141
$\pi_1 = 1\%$	500	0.980	1.617	1.653	1.519	1.519	1.124

Table 2.3: The bias of the new method and the standard method (only controls) in the heteroscedastic simulation of Section 2.5.2 for $n_0 = 100, 300$ and 500.

		New Method				Controls only			
	X_0	0.19	0.39	0.59	0.79	0.19	0.39	0.59	0.79
	100	-0.007	-0.041	-0.062	-0.054	-0.016	-0.049	-0.063	-0.045
$\pi_1 = 5\%$	300	-0.01	-0.03	-0.051	-0.053	-0.019	-0.035	-0.049	-0.044
	500	-0.011	-0.027	-0.049	-0.052	-0.019	-0.031	-0.048	-0.044
	100	-0.003	-0.032	-0.048	-0.037	-0.011	-0.04	-0.048	-0.027
$\pi_1 = 1\%$	300	-0.007	-0.023	-0.039	-0.033	-0.015	-0.029	-0.038	-0.024
	500	-0.008	-0.021	-0.034	-0.034	-0.016	-0.025	-0.033	-0.025

Table 2.4: The mean square efficiency of the new method and the standard method (only controls) in the heteroscedastic simulation of Section 2.5.2 for $n_0 = 100, 300$ and 500.

		MSE Efficiency					
	X_0	0.01	0.19	0.39	0.59	0.79	0.99
$\pi_1 = 5\%$	100	1.187	1.401	1.446	1.498	1.371	1.432
	300	1.075	1.490	1.439	1.357	1.254	1.198
	500	1.042	1.437	1.456	1.274	1.178	1.169
$\pi_1 = 1\%$	100	1.055	1.355	1.637	1.499	1.499	1.459
	300	1.010	1.292	1.546	1.335	1.271	1.295
	500	0.983	1.361	1.540	1.208	1.183	1.266

2.6 Empirical Example

2.6.1 Colorectal Adenoma Study

The rate of occurrence of colorectal adenoma (D), responsible for the colon cancer is typically not known. So we implemented our "tilted efficient" method on a case-control data set with 640 cases and 665 controls. The cases and controls were defined by the occurrence of colorectal adenoma (D). In our analysis, X is red meat consumption in grams. We used two different versions of Y, namely the heterocyclic amines MeIQx and PhIP (both measured in nanograms) that are produced during the cooking of meat. PhIP and MeIQx were transformed by adding 1.0 and taking logarithms, while red meat was transformed as 10 plus the logarithm. We calculated the mean square errors by the bootstrap method, with 1000 bootstrap samples.

Preliminary analysis of the controls data indicated a highly statistically significant linear effects of red meat consumption on MEIQx and PHIP. The p-values for the coefficient for a quadratic fit exceeds 0.20 in both the cases. In addition, the regression of PHIP on red meat consumption is heavily heteroscedastic, while the regression of MeIQx on red meat is passably homoscedastic. For MeIQx, where the regression is homoscedastic, both the controls and new method have roughly the same bias. In this case however, the new method achieves an approximately 150% average efficiency gain in efficiency as expected. For PhIP, where the regression is heteroscedastic, the MSE average efficiency gain of the new method is almost 150%, see Figure 2.3.

2.6.2 NIH-AARP Study

As a further illustration, we used data from the National Institutes of Health-AARP Diet and Health Study (NIH-AARP). We constructed case-control studies from these data with 4 controls for every case. Here the outcome D was incidence of

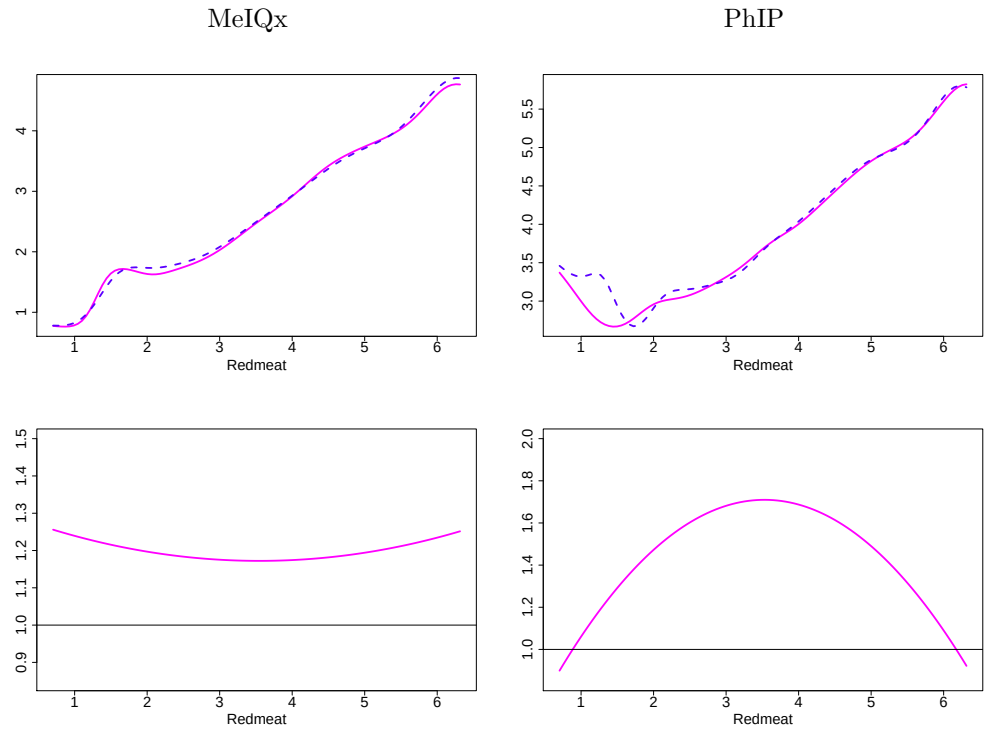

Figure 2.3: Results for the data analysis in Section 2.6.1. First row are the fitted functions in the kernel regression of MeIQx and PhIP on red meat, respectively, with the magenta solid line being our method and the blue dashed line using the controls only. The second row are the pointwise mean squared error efficiency of our method for the two responses (solid magenta line). The dashed black line is at 1.0, with values above that indicating that our method is more efficient

colorectal cancer, responsible for colon cancer. We did separate analysis for men and women, in the latter case deleting those with missing menopausal hormone therapy status, none of whom developed colorectal cancer. In this study, the sample sizes were $n = 21240$ with 4248 number of cases and 16992 number of controls. We performed separate studies of the association between body mass index (BMI) and Fat Density as Y, and Age as X. In the second study we regress Y on Alcohol intake as X. We did a preliminary analysis among the controls and found out a strong quadratic relationship exist in the four different associations with p-value less than $2e^{-16}$. Also, the regression of BMI and Fat Density is heavily heteroscedastic

on alcohol intake while that on the age is highly homoscedastic. We calculated the mean square errors by the bootstrap method, with 500 bootstrap samples.

In both the study the estimates obtained from "tilted efficient" method (2.10) and only controls method has almost same estimates except for regressing BMI with respect to age. In all the scenarios our estimator shows a clear quadratic trend. The MSE average efficiency gain of the new method is approximately 110% for the homoscedastic regression on age. In case of heteroscedastic regression on alcohol intake the average MSE efficiency gain is almost 115%, see Figure 2.4.

2.7 Discussion

Primarily, we have considered the case of nonparametrically estimating $E(Y|X)$ when no assumptions about the distribution of Y given X are made, including homoscedasticity. Section 3.2.2 describes methodology in the rare case that the disease rate in the population, π_1, is known. Sections B.1 - 3.2.2 describe methodology in the far more common case that π_1 is unknown. In this common case, our simulations show conclusively that our tilted kernel estimator defined in Section 3.2.2 is the more efficient.

In Section 2.4, we considered the case that the disease rate in the population is unknown, and when one is willing to specify a distribution for Y given X up to a function $\mu(X)$ and other parameters, using a local likelihood method along with profiling methods. We displayed the method for when Y is binary with mean $H\{\mu(X)\}$. However, we emphasized two important points: (a) such methods are not consistent if the parametric model is misspecified; and (b) it is likely that the logistic intercept θ_0 will be very difficult to estimate numerically, and a rare disease approximation will improve computational performance (since it eliminates θ_0) while entailing little if any bias.

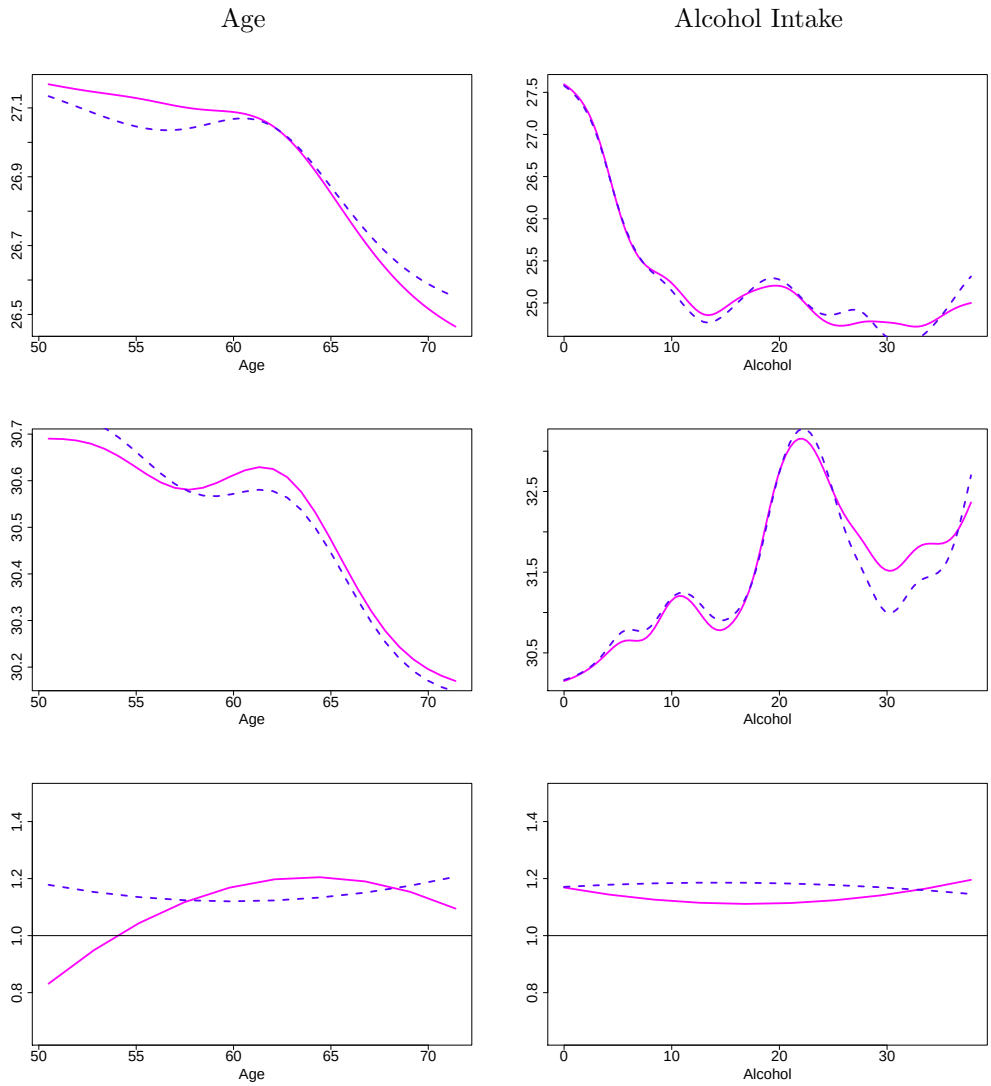

Figure 2.4: Results for the data analysis in Section 2.6.2. First row are the fitted functions in the kernel regression of BMI on age and alcohol content, respectively, with the magenta solid line being our method and the blue dashed line using the controls only. The second row are the fitted functions of Fat-density on the same regressors. The third row are the pointwise mean squared error efficiencies of our method for the two responses (solid magenta line for BMI and dashed blue line for Fat-density). The dashed black line is at 1.0, with values above that indicating that our method is more efficient

Ours is the first paper to consider nonparametric regression in the secondary analysis of case-control studies. We have focused on the case of scalar X, and discovered a

tilted kernel approach for estimation. With this tilted kernel function, extensions to multivariate X are surely possible, including multivariate kernel regression (Ruppert and Wand, 1994), additive models, etc.

3. BAYESIAN SINGLE-INDEX MODELING WITH VARIANCE ESTIMATION

3.1 Single-Index Model with Variance Function

We consider the heteroscedastic regression models where the mean function is a single-index model and the variance function depends on another single-index model, so that for $i = 1, ..., n$,

$$Y_i = m(\boldsymbol{X_i}^{\mathrm{T}}\boldsymbol{\theta}) + s(\boldsymbol{X_i}^{\mathrm{T}}\boldsymbol{\gamma})\epsilon_i, \tag{3.1}$$

where the Y_i are scalar continuous response variables, the $\boldsymbol{X_i}$ are p-dimensional continuous predictors and ϵ is independent of $\mathbf{X}$. The unknowns are the p-dimensional index vectors $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, the regression function $m(\cdot)$ and standard deviation function $s(\cdot)$. The regression errors ϵ are independent and identically distributed according to some density $f_\epsilon(\cdot)$, with restriction $E(\epsilon) = 0$ and $E(\epsilon^2) = 1$ to ensure identifiability. Hence, the conditional heteroscedasticity is only explained through the variance function $s^2(\cdot)$. To avoid identifiability issues, we need additional restrictions on the single-index vectors, specifically, $\|\boldsymbol{\theta}\| = \|\boldsymbol{\gamma}\| = 1$.

To develop an efficient and practical Bayesian estimation methodology for the index-vectors $\boldsymbol{\theta}, \boldsymbol{\gamma}$, the mean function $m(\boldsymbol{X_i}^{\mathrm{T}}\boldsymbol{\theta})$ and the variance function $s^2(\boldsymbol{X_i}^{\mathrm{T}}\boldsymbol{\gamma})$, we consider the following MCMC estimation procedure. Our development proceeds in the following step.

(a) Initial estimation for the starting values for the index parameter $\boldsymbol{\theta}$.

(b) Using the squared residuals from the initial estimates of the mean function, develop an initial estimate of the index vector $\boldsymbol{\gamma}$.

(c) Design a non-informative prior for each of the parameters.

(d) Update the parameters for the mean function, including the index vector $\boldsymbol{\theta}$ in each MCMC iteration step.

(e) Update the parameters for the variance function, including $\boldsymbol{\gamma}$ in each MCMC step.

In this paper we use the definition of the inverse-gamma and gamma distribution from Berger (1985),

$$\mathrm{IG}(A, B) = \frac{1}{\Gamma(A) B^A x^{A+1}} \exp\Big\{ -1/(Bx) \Big\} I_{(0,\infty)}(x)$$

and

$$\mathrm{Gamma}(A, B) = \frac{1}{\Gamma(A) B^A} x^{A-1} \exp\Big\{ -x/B \Big\} I_{(0,\infty)}(x).$$

The generic notation p_0 is used for specifying priors and hyperparameters. The notation $\mathrm{Normal}(\cdot|\mu, \sigma^2)$ is used to denote a normal distribution with mean μ and variance σ^2.

3.1.1 Estimation of the Single-Index Vectors

In this problem, the mean and the variance functions are depend on the single-index vectors, θ and γ, and are thus crucial in the mean and variance estimation. Essentially, the single index parameter helps in the reduction of dimension and boils down to estimate the central mean space $E(Y|\boldsymbol{X})$ by $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}$ (Cook and Li, 2002) and $E(\epsilon^2|\boldsymbol{X})$ by $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma}$. In this article, our goal is to introduce a fully Bayesian method to estimate the index vectors. If we want to use a non-informative prior, then selecting the initial value for the single-index vector is crucial. To make the estimation

procedure simple, we will use the Principal Hessian Direction (PHD) method (Li, 1992), only to decide on the starting value for the single-index parameter.

Initial values for $\boldsymbol{\theta}$ *and* $\boldsymbol{\gamma}$: Define $\Lambda_{\text{PHD}} = E[\{Y - m(X^{\text{T}}\boldsymbol{\theta})\}^2 XX^{\text{T}}]$ and $\Gamma_{\text{PHD}} = E[\{|\epsilon| - s(X^{\text{T}}\boldsymbol{\gamma})\}^2 XX^{\text{T}}]$. This method is simple and requires only computing the eigenvector associated with the maximum non-zero eigenvalue of the Principal Hessian matrices Λ_{PHD} and Γ_{PHD} to form the bases of the subspace, $\mathcal{S}_{E(Y|X)}$ and $\mathcal{S}_{E(\epsilon^2|X)}$ respectively. We define $\overline{Y} = n^{-1}\sum_{i=1}^{n} Y_i$, $e_i = |Y_i - \overline{Y}|$ and $\overline{e} = n^{-1}\sum_{i=1}^{n} e_i$. To get a reasonable starting value, we estimate Λ_{PHD} by $\widehat{\Lambda} = n^{-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 X_i X_i^{\text{T}}$ and Γ_{PHD} by $\widehat{\Gamma} = n^{-1}\sum_{i=1}^{n}(e_i - \overline{e})^2 X_i X_i^{\text{T}}$. To make $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ identifiable, we can either set the norm of the vectors to 1 or fix one of the component at 1. Without loss of generality, we fix the first component of $\boldsymbol{\theta}$ to 1, defining the other components of $\boldsymbol{\theta}$ as $\boldsymbol{\theta}_{-1} = (\theta_2, ..., \theta_p)$. We use the eigenvector corresponding to the maximum eigenvalue of $\widehat{\Lambda}$ as a *starting value* for $\boldsymbol{\theta}_{-1}$. Similarly we denote $\boldsymbol{\gamma}_{-1} = (\gamma_2, ..., \gamma_p)$ and calculate the *starting value* for $\boldsymbol{\gamma}_{-1}$ using the maximum eigen value of $\widehat{\Gamma}$. For each MCMC iteration step, we fix $\theta_1 = \gamma_1 = 1$ and specify normal priors on the rest of the components of the vectors

$$p_0(\boldsymbol{\theta}_{-1}) = \text{Normal}(\cdot|\boldsymbol{\theta}_{\text{prior}}, \Sigma_{\boldsymbol{\theta}}), \tag{3.2}$$

$$p_0(\boldsymbol{\gamma}_{-1}) = \text{Normal}(\cdot|\boldsymbol{\gamma}_{\text{prior}}, \Sigma_{\boldsymbol{\gamma}}), \tag{3.3}$$

where $\boldsymbol{\theta}_{\text{prior}}, \boldsymbol{\gamma}_{\text{prior}}$, $\Sigma_{\boldsymbol{\theta}}$ and $\Sigma_{\boldsymbol{\gamma}}$ are pre-specified constants.

3.1.2 Estimation of the Mean Function

If the potential heteroscedasticity is ignored, methods for estimating the flexible and smooth mean function using penalized splines are already well established (Yu

and Ruppert, 2002; Antoniadis, et al., 2004). Given $\boldsymbol{\theta}$, a general approach discussed by Eilers and Marx (1996) is to model the regression function $m(\cdot)$ by a linear combination of *basis splines* or *B-splines* with fixed knots and smoothing parameter ρ_1, better known as *penalized splines* or *P-splines*. Let $B_{1j}(t)$ be the j^{th} B-spline, recursively defined by De Boor (2001) of order q for the knot sequence $T = t_{j=1}^{K}$ is a sequence of equally-spaced points, called *interior knots*. Augment these so that $t_{1-q} = \cdots = t_{-1} = t_0 = 0 < t_1 < \cdots < t_K < 1 = t_{K+1} = \cdots = t_{K+q}$, in which $t_j = jh$ for $j = 0, \cdots, K+1, h = 1/(K+1)$ is the distance between neighboring knots. Define $\mathcal{B}_1(t) = \{B_{1j}(t)\}_{j=1}^{M}$, where $M = K + q - 2$. The mean function is evaluated at the "transformed design points" $t_i = \boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\theta}}$ for $i = 1, ..., n$ as

$$\widehat{m}(t_i) = \sum_{j=1}^{M} \mathcal{B}_{1j}(t_i)\beta_j. \tag{3.4}$$

Let D be a fixed, symmetric and a positive semidefinite N-dimensional matrix. The penalized least squares estimator $\widehat{\boldsymbol{\beta}}(\rho_1, \boldsymbol{\theta})$ for the mean function minimizes

$$\sum_{i=1}^{n} \{Y_i - \mathcal{B}_1(t_i)^{\mathrm{T}}\boldsymbol{\beta}\}^2 + \rho_1 \boldsymbol{\beta}^{\mathrm{T}} D \boldsymbol{\beta}.$$

Here $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_M)^{\mathrm{T}}$ and $D = D_r^{\mathrm{T}} D_r$ is the matrix representation of the difference operator Δ_r of order r defined by Eilers and Marx (1996). For simplicity of notation, we define $\widetilde{\mathbf{X}}_{\boldsymbol{\theta}} = \mathcal{B}_1(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})$ to obtain the least square estimators for $\boldsymbol{\beta}$ as

$$\widehat{\boldsymbol{\beta}}(\rho_1, \boldsymbol{\theta})_{LS} = (\widetilde{\mathbf{X}}_{\boldsymbol{\theta}}^{\mathrm{T}}\widetilde{\mathbf{X}}_{\boldsymbol{\theta}} + \rho_1 D)^{-1}\widetilde{\mathbf{X}}_{\boldsymbol{\theta}}^{\mathrm{T}} Y.$$

We can use $\widehat{\boldsymbol{\beta}}_{LS}$ as a potential *starting value* for the MCMC iteration. We model the mean function by a B-spline with smoothness inducing prior on the coefficients,

given as

$$\begin{aligned} p_0(\boldsymbol{\beta}) &\propto \rho_1^{M/2} \exp\{-\rho_1 \boldsymbol{\beta}^{\mathrm{T}} D \boldsymbol{\beta}/2\}, \\ \rho_1 &\sim \text{Gamma}(A_{\rho 1}, B_{\rho 1}). \end{aligned} \tag{3.5}$$

The prior $p_0(\boldsymbol{\beta})$ induces smoothness in the coefficients because it penalizes $\sum_{j=r+1}^{n} (\Delta_r \beta_j)^2 = \boldsymbol{\beta}^{\mathrm{T}} D \boldsymbol{\beta}$, the sum of squares of the second order difference in $\boldsymbol{\beta}$. Determining a suitable smoothing parameter in spline methods is crucial. Antoniadis et al. (2004) provided a semi-Bayesian method to model the mean function for single-index model by P-splines where the smoothing parameter was determined by generalized cross validation (GCV). In that case, estimation of the smoothing parameter would depend on the index-vectors, making the problem computationally difficult. However, in a fully Bayesian approach we place a continuous prior distribution on the smoothing parameter. That automatically avoids the possibility of zero smoothing and the estimation of the smoothness parameter becomes independent of the index-vectors.

3.1.3 Estimation of the Variance Function

We define the true absolute residual as $|\epsilon_i| = |Y_i - m(\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\theta})|$ for $i = 1, \ldots, n$. In the parametric case, Davidian and Carroll (1987) gave the general methodology and theory for the variance function. They pointed out that estimation of variance function based on squared residuals are less robust to outliers than those based on absolute residuals. Lian et al. (2013) used two-stage approach to model the variance function by estimated absolute residuals $\widehat{r} = |Y - \widehat{m}(\boldsymbol{X}^{\mathrm{T}} \widehat{\boldsymbol{\theta}})|$. In our Bayesian approach, estimation of a flexible variance function $s(\cdot)$ and the corresponding single-index vector $\boldsymbol{\gamma}$ is straightforward. We only need to specify the prior distribution of

the parameters. Examples of modeling log-transformed variance functions by flexible mixture of splines are abundant in the literature, for example Yau and Kohn (2003), Liu, et al. (2006). In this article, for fixed $\boldsymbol{\gamma}$, we model the variance function by a positive mixture of B-splines of order q with K_1 knots at transformed points $w_i = \boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}$ as

$$s(w_i) = \sum_{j=1}^{M_1} \mathcal{B}_{2j}(w_i)\exp(\xi_j) \tag{3.6}$$

where, $M_1 = K_1+q-2$. For simplicity, we define $\widetilde{\mathbf{X}}_{\boldsymbol{\gamma}} = \mathcal{B}_2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})$. Then a reasonable starting value for the P-spline coefficient $\boldsymbol{\xi}$ of the variance function is

$$\exp\left\{\widehat{\boldsymbol{\xi}}(\rho_2,\boldsymbol{\gamma})_{LS}\right\} = |(\widetilde{\mathbf{X}}_{\boldsymbol{\gamma}}^{\mathrm{T}}\widetilde{\mathbf{X}}_{\boldsymbol{\gamma}} + \rho_2 D)^{-1}\widetilde{\mathbf{X}}_{\boldsymbol{\gamma}}^{\mathrm{T}}\widehat{r}^2|,$$

A flexible Bayesian model for the variance function with smoothness inducing priors on the coefficients is

$$\begin{aligned} p_0(\boldsymbol{\xi}) &\propto \rho_2^{M_1/2}\exp\{-\rho_2\exp(\boldsymbol{\xi})^{\mathrm{T}} D\exp(\boldsymbol{\xi})/2\}, \\ \rho_2 &\sim \mathrm{Gamma}(A_{\rho 2}, B_{\rho 2}). \end{aligned} \tag{3.7}$$

Here ρ_2 is the smoothing parameter for the variance function. Larger values of ρ_2 with smaller number of knots K_1 imposes a stronger penalty, resulting in smoother variance function.

In this paper, the predictor variables $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}$ and $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma}$ changes with each MCMC iteration of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$. We fix K_1 and K_2, the numbers of knots to model the mean function and the variance function respectively. In each iteration, we place the knots equally spaced in $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}$ and $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma}$. For smooth and either monotonic or unimodal

regression functions, Ruppert (2000), Yu and Ruppert (2002) recommended 5-10 knots to be adequate. However, more than 10 knots are needed if the true function has many local minima and maxima, but this is unlikely in applications of the single-index model. Yu and Ruppert (2002) also showed that fixed-knot asymptotics give a practical result, converging to a known normal distribution.

3.2 Estimation of the Density Function of the Error

We consider f_ϵ as the density of the errors for the estimation of the mean function $m(\cdot)$. In this paper we consider two different distributions in modeling the density of the scaled errors, (I) standard normal distribution and (II) DPMM. While estimation based on (I) is straightforward, simple and produces consistent results, method (II) can capture multimodality and heavy tails. By using (II), along with consistency, we can achieve a potential gain in efficiency in small or moderate sample size.

3.2.1 Model-I: Normal Distribution

In this method we take a standard normal distribution for modeling the density of ϵ

$$f_\epsilon(\epsilon) = \text{Normal}(\epsilon|0,1). \tag{3.8}$$

This implies that the conditional distribution of Y given $\boldsymbol{X}$ is Normal$\{\cdot|m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}),$ $s^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\}$.

3.2.2 Model-II: Dirichlet Process Mixture Models (DPMM)

Misspecification of the error distribution may lead to inefficient estimation, especially for heavy tailed distribution. The model can be robustified by assuming that the error ϵ is modeled nonparametrically. To do so, in recent Bayesian literature, there has been an explosion of interest in the Bayesian nonparametric methods due

to their flexibility and to the availability to the easy to use algorithms for posterior computations. Most of the focus has been on modeling the distribution of the error ϵ by DPMM. See for example, Bush and MacEachern (1996), Gelfand et al. (2005), Leslie et al. (2007), Griffin and Steel (2010), Pelenis (2014).

For modeling a density $f(\cdot)$, a DPMM, usually denoted as $\mathrm{DP}(\alpha P_0)$ with concentration parameter α, base measure P_0 and mixture components coming from a parametric family $\{f_c(\cdot|\phi) : \phi \sim P_0\}$, can be specified as

$$f(\cdot) = \sum_{k=1}^{\infty} \pi_k f_c(\cdot|\phi_k), \quad \phi_k \sim P_0, \quad \pi_k = \pi_k^* \prod_{i=1}^{k-1}(1-\pi_j^*), \quad \pi_k^* \sim \mathrm{Beta}(1,\alpha),$$

In the literature, this construction of random mixture weights $\{\pi_k\}_{k=1}^{\infty}$, illustrated first by Sethuraman (1994), is known as Stick-breaking representation, hence $\pi \sim \mathrm{Stick}(\alpha)$. Lo (1984) showed that a DPMM of normal density, that is $f_c(\cdot) = \mathrm{Normal}(\cdot)$ is dense in the space of densities with respect to Lebesgue measure. Hence, DPMMs of normals are popular for modeling densities (Escobar and West, 1995; West, et al. 1994). In the context of regression analysis, moment constraint infinite mixture models of Normal has been considered by Griffin and Steel (2010) and Pelenis (2014), so that the mean of the error can be restricted at 0.

Drawing inspiration from them, we let $f_c(\cdot)$ to be a two component mixture of normal as

$$f_c(\cdot|p,\mu_1,\mu_2,\sigma_1^2,\sigma_2^2) = p\mathrm{Normal}(\cdot|\mu_1,\sigma_1^2) + (1-p)\mathrm{Normal}(\cdot|\mu_2,\sigma_2^2),$$

subject to the moment constraint

$$p\mu_1 + (1-p)\mu_2 = 0. \tag{3.9}$$

We define $\mu_k^* = -p\mu_k/(1-p)$ to obtain the complete DPMM prior for the density function of ϵ, namely,

$$\begin{aligned} f_\epsilon(\epsilon) &= \sum_{k=1}^{\infty} \pi_k f_c(\epsilon | p_k, \mu_k, \mu_k^*, \sigma_{k1}^2, \sigma_{k2}^2), \\ \boldsymbol{\pi} &\sim \text{Stick}(\alpha), \qquad p_k \sim \text{Unif}(0,1), \\ \mu_k &\sim \text{Normal}(a_0, b_0), \qquad \sigma_{k1}^2, \sigma_{k2}^2 \sim \text{IG}(c_0, d_0). \end{aligned} \tag{3.10}$$

Here α, b_0, c_0 and d_0 are positive preassigned constants.

Ishwaran and James (2001) constructed an useful class of $\text{DP}_C(\alpha P_0)$ process which are constructed by applying a truncation to $\text{DP}(\alpha P_0)$ process. The truncation is applied by discarding $C+1, C+2, \cdots$ terms in the infinite $\text{DP}(\alpha P_0)$ process and replacing π_C with $1 - p_1 - \cdots - p_{C-1}$. Determination of appropriate truncation level can be based on the moments of the random weights. They also showed that for each positive integer $r \geq 1$, $\text{E}(\sum_{k=C}^{\infty} p_k)^r$ and $\text{E}(\sum_{k=C}^{\infty} p_k^r)$ decreases exponentially fast in C and, thus, for a moderate C, we should be able to achieve an accurate approximation. They have also given precise bound for marginal density of Y, g_C under $\text{DP}_C(\alpha P_0)$ process, $\|g_C - g_\infty\|_1 \sim 4n \exp\{-(N-1)/\alpha\}$, where $\|\cdot\|_1$ denotes the $\mathcal{L}_1$ distance and g_∞ denotes the marginal density of Y under $\text{DP}(\alpha P_0)$. Hence for $\alpha = 1$, $C = 20$ and $C = 10$ we get an $\mathcal{L}_1$ bound of order 10^{-6} and 10^{-3} respectively. Therefore, even for a sample size of 500, a mere truncation of $C = 10$ leads to an approximating hierarchical model that is virtually indistinguishable from one based on the $\text{DP}(\alpha P_0)$ prior. See Ishwaran and James (2000) for more discussion and for application of this truncation to estimate finite mixture of normals.

Thus we consider truncated $DP(\alpha P_0)$ process of finite mixture of densities upto

a fixed number C as follows

$$f(\cdot) = \sum_{k=1}^{C} \pi_k f_c(\cdot|\phi_k), \quad \phi_k \sim P_0, \quad \pi_k = \pi_k^* \prod_{i=1}^{k-1}(1-\pi_j^*), \quad \pi_k^* \sim \text{Beta}(1,\alpha),$$

Blocked Gibbs Sampler can be applied to finite dimensional DP(αP_0). The finite dimensionality of such priors is a key to the success of the method because it allows us to express our model entirely in terms of a finite number random variables. This then allows the blocked Gibbs sampler to update (Block of parameters), which because of the nature of the prior, are drawn from simple multivariate distributions. Other screening techniques by Berkhof et al. (2003) and Li and Chen (2010) for number of components for the finite mixture model can also be used to determine C.

3.2.3 Identifiability of Standard Deviation Function

In section 3.1, we define the single-index model in equation (3.1) where we put a restriction on the second order moment of error to identify the standard deviation function $s(\cdot)$. However in the previous section we did not put any such restrictions on the distribution of the error. If we impose the restriction on the distribution on error, we need to put severe constraints on the prior of the location and scale parameters on the error. To avoid such a complicate approach, we implement an extremely simple way to follow the constraint. All we need to do is to adjust a positive constant ($0 < a < \infty$) in the error and the standard deviation function as follows.

$$\begin{aligned} s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\epsilon &= \frac{s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})}{a} \times a\epsilon \\ &= \widetilde{s}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\widetilde{\epsilon} \end{aligned}$$

Here, $\widetilde{s}(\cdot) = s(\cdot)/a$ and $\widetilde{\epsilon} = \epsilon/a$. Hence $\widetilde{\epsilon}$ becomes free of second order moment restriction since $\mathrm{E}(\widetilde{\epsilon}^2) = a^2 > 0$. Thus $\widetilde{\epsilon}$ could be modeled efficiently by the penalized method discussed in section 3.1.3. Then the original standard deviation function $\widehat{s}(\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}})$ can be estimated simply by $\widehat{a} \times \widetilde{s}(\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}})$. The constant $\widehat{a}$ is consistently estimated by the standard deviation of the estimated distribution of $\widetilde{\epsilon}$. That is, $\widehat{a}^2 = \sum_{k=1}^{C} \widehat{\pi}_k\{p_k(\widehat{s}_{k1}^2 + \widehat{\mu}_{k1}^2) + (1-p_k)(\widehat{s}_{k2}^2 + \widehat{\mu}_{k2}^2)\}$, where $p_k\mu_{k1} + (1-p_k)\mu_{k2} = 0$.

3.3 Simulation Studies

3.3.1 Basic Settings

In this section we conduct a series of simulations based on different sample sizes $n = 200, 500$ and 1000 to compare the kernel method of Lian, et al. (2014) with our normal model (Model-I) and the DPMM model (Model-II). Data are generated from a "sine-bump" model, a design similar to that of Carroll et al. (1997), namely,

$$y_i = \sin\big\{\ \pi(X_i^{\mathrm{T}}\boldsymbol{\theta} - A)/(B-A)\big\} + \big\{0.2 + (X_i^{\mathrm{T}}\boldsymbol{\gamma})^2/8\big\}\epsilon_i, \tag{3.11}$$

where $X = (X_1, X_2, ..., X_8)^{\mathrm{T}}$, with each component being independent Uniform$(0,1)$ with $A = \sqrt{3}/2 - 1.645/\sqrt{12}$ and $B = \sqrt{3}/2 + 1.645/\sqrt{12}$.

We take $\boldsymbol{\theta}^{\mathrm{T}} = (1,1,1,1,0.5,0.5,-0.5,-0.5)/\sqrt{5}$ and $\boldsymbol{\gamma}^{\mathrm{T}} = (1,1,1,1,0.5,0.5,0.5,0.5)$ $/\sqrt{5}$. Including a burn-in of 3000 iterations, we use a total of $10,000$ MCMC iterations to obtain the posterior average of each parameters. In each case, 100 simulated datasets were generated with 3 different error models, namely, (I) normal (mean = 0, $\sigma = 1$), (II) Scaled Gamma (mean = 0, scale = 1), and (III) Mixture of Normals (mean = 0, scale = 1). We use P-splines of order 3 with 10-knots for the mean function and use 5 knots for variance function estimation.

In the simulation, for the fully Bayesian method, the following prior distributions

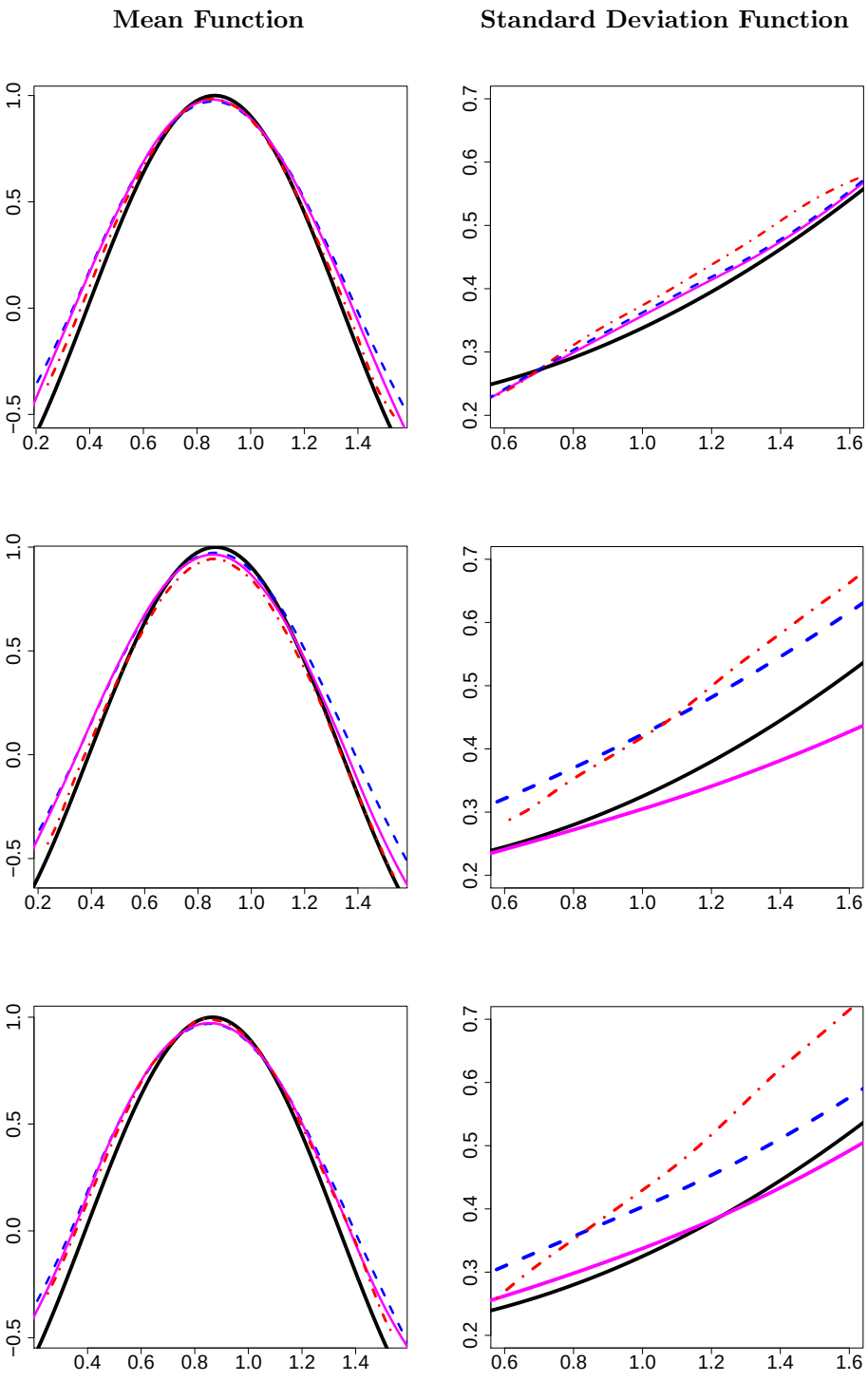

Figure 3.1: The mean and standard deviation function estimation for 3 different error distribution, (a) Normal (first row), (b) scaled Gamma (second row) and (c) Mixture of Normals (third row) based on sample size 500. The "blue dashed curve" is the estimate from Method-I (Normal errors). The "solid magenta curve" and "red dot-dashed curve" represents DPMM and local linear kernel method (LLK1) respectively.

are used

1. *Index vectors:* θ_i, $\gamma_i \sim \text{Normal}(0, 10)$ for $i = 2, ..., 8$.

2. *Smoothing parameter:* ρ_1, $\rho_2 \sim \text{IG}(3, 1)$.

3. *B-spline coefficients:* $p_0(\boldsymbol{\beta}) \propto \rho_1 \boldsymbol{\beta}^{\mathrm{T}} D_r^{\mathrm{T}} D_r \boldsymbol{\beta}$ and $p_0(\boldsymbol{\xi}) \propto \rho_2 \exp(\boldsymbol{\xi})^{\mathrm{T}} D_r^{\mathrm{T}} D_r \exp(\boldsymbol{\beta})$. The matrix D_r can be computed by using a built-in function `diff()` of order r on an identity matrix in `R`.

4. *DPMM model parameters:* We take a finite number of clusters $C = 10$; $\boldsymbol{\pi} \sim$ Stick(1); $\mu_k \sim \text{Normal}(0, 1)$; $\sigma_{k1}^2, \sigma_{k2}^2 \sim \text{IG}(1, 1)$ for $k = 1, ..., 10$.

The priors are all proper but not informative. We found the results insensitive to moderate modifications of these priors. For purpose of bias and mean squared error calculations, the P-spline estimates of mean and variance functions were computed on a grid of 100 points in the interval chosen to contain the distribution of X. In the Metropolis-Hasting steps we choose the proposed distributions to be symmetric. The choice of the "width" of the proposed distribution is sometimes crucial in achieving a good mixing of the Markov chains. We define $\mathcal{I}$ as a diagonal matrix, $\text{Uniform}_+(a - \delta, a + \delta)$ as the absolute value from $\text{Uniform}(a - \delta, a + \delta)$ if the random value is less than 0. For sampling the probability weights $\boldsymbol{\pi}$, if the value is greater than 1, then the distribution returns $2 - \text{Uniform}(a - \delta, a + \delta)$. The proposal distributions for the parameters of the MCMC chain are as follows.

i. *Index vectors:* $\boldsymbol{\theta}_{-1} \sim \text{Normal}\{\boldsymbol{\theta}_{-1(n)}, 0.01\mathcal{I}\}$; $\boldsymbol{\gamma}_{-1} \sim \text{Normal}\{\boldsymbol{\gamma}_{-1(n)}, 0.01\mathcal{I}\}$.

ii. *B-spline coefficients:* $\boldsymbol{\beta} \sim \text{Normal}\{\boldsymbol{\beta}_{(n)}, 0.001\mathcal{I}\}$; $\boldsymbol{\xi} \sim \text{Normal}\{\boldsymbol{\xi}_{(n)}, 0.01\mathcal{I}\}$.

iii. *DPMM model parameters:* $\boldsymbol{\pi} \sim \text{Uniform}_+\{\boldsymbol{\pi}_{(n)} - 0.01, \boldsymbol{\pi}_{(n)} + 0.01)\}$;
$\boldsymbol{\mu} \sim \text{Normal}\{\boldsymbol{\mu}_{(n)}, 0.001\mathcal{I}\}$; $\boldsymbol{\sigma}_1^2 \sim \text{Uniform}_+\{\boldsymbol{\sigma}_{1(n)}^2 - 0.1, \boldsymbol{\sigma}_{1(n)}^2 + 0.1\}$;

$$\boldsymbol{\sigma}_2^2 \sim \text{Uniform}_+\{\boldsymbol{\sigma}_{2(n)}^2 - 0.1, \boldsymbol{\sigma}_{2(n)}^2 + 0.1\}.$$

For each $d = 1, \cdots, 100$ data set, let $\widehat{\boldsymbol{\theta}}_d$ be the posterior estimate of the index vector $\boldsymbol{\theta}$. The Monte Carlo estimate of root mean squared error (RMSE) for $\boldsymbol{\theta}$ is calculated as $S^{-1}\sum_{d=1}^{S}\|\widehat{\boldsymbol{\theta}}_d - \boldsymbol{\theta}\|/p$. The RMSE for the index vector $\boldsymbol{\gamma}$ is calculated similarly. Let $X_{\text{grid}} = (X_1, \cdots, X_{100})$ be equally-spaced 100 grid points on an interval $[0, 1]$. RMSE for the smooth function $m(\cdot)$ at $\boldsymbol{Z}_{\text{grid}} = \boldsymbol{X}_{\mathbf{grid}}^{\mathrm{T}}\theta$ is computed as $\text{MISE}_{est} = (100S)^{-1}\sum_{d=1}^{S}\sum_{i=1}^{100}\|\widehat{m}(\boldsymbol{Z}_{\text{grid}}) - m(\boldsymbol{Z}_{\text{grid}})\|$, where $\|\cdot\|$ is the euclidean distance for norm of a vector. The RMSE for the index vector $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are given in Table 3.1. In Table 3.2, we calculate the RMSE for the mean and standard deviation functions for each settings. We calculate the *Relative Efficiency* of Method II over Method I as $p^{-1}\sum_{i=1}^{p}$ (RMSE of Method II for θ_i/ RMSE of Method I for θ_i). Similarly we compare the efficiency of Method II over Kernel method for estimating θ and γ.

We show in the simulations that the method based on local linear kernel (LLK) method of Lian et al. (2014) is sensitive to the initial values of the single-index parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$. We take two different set up denoted as LLK1 and LLK2 to compare the performances. In LLK1 we choose the starting value close to the truth while in LLK2 we chose estimates from the Principal Hessian Direction method (PHD) which is used for our methods. In Table 3.1, we show that the estimates of the single-index vector looses huge amount of efficiency for LLK2, thus selection of good starting value is important for the local linear method to yield consistent estimates. For each parameter, the "Average Efficiency" is calculated for the robust method (DPMM) with respect to other methods, where the average is taken over the three different sample sizes. When the true distribution of error is normal, then our method based on normal likelihood (Normal) yields minimum root mean square

	$\boldsymbol{\theta}$				$\boldsymbol{\gamma}$			
	LLK	Normal	DPMM	LLK2	LLK	Normal	DPMM	LLK2
$\epsilon \sim$ Normal								
$n = 200$	0.049	0.045	0.051	0.168	0.234	0.211	0.226	0.153
$n = 500$	0.033	0.025	0.027	0.144	0.149	0.137	0.140	0.152
$n = 1000$	0.021	0.016	0.016	0.135	0.099	0.096	0.101	0.164
Avg. Efficiency	**1.159**	**0.928**		**5.689**	**1.031**	**0.958**		**1.129**
$\epsilon \sim$ Gamma								
$n = 200$	0.047	0.043	0.038	0.176	0.148	0.179	0.198	0.149
$n = 500$	0.030	0.025	0.018	0.139	0.122	0.148	0.145	0.154
$n = 1000$	0.022	0.017	0.011	0.129	0.106	0.124	0.092	0.168
Avg. Efficiency	**1.596**	**1.326**		**8.209**	**0.914**	**1.091**		**1.199**
$\epsilon \sim$ Mixture of Normals								
$n = 200$	0.051	0.044	0.023	0.177	0.114	0.167	0.162	0.154
$n = 500$	0.031	0.025	0.011	0.137	0.105	0.134	0.077	0.158
$n = 1000$	0.020	0.017	0.007	0.133	0.100	0.104	0.042	0.171
Avg. Efficiency	**2.597**	**2.172**		**13.05**	**1.498**	**1.757**		**2.358**

Table 3.1: The table shows the "root mean squared error" for single-index vectors $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, which is evaluated for each of the 100 datasets under 3 different error distributions, namely, (1) Normal $(0, 1)$, (2) Scaled Gamma (mean = 0, sd = 1) and (3) Mixture of normals (mean = 0, scale = 1). For each error distribution, we compare the performance of the new methods,(I) Normal and (II) DPMM with the local linear kernel method (LLK, when the starting value is closer to true value and LLK2, when the starting values are chosen by Principal Hessian Direction Method) for samples sizes, $n = 200$, 500 and 1000. RMSE for $\widehat{\boldsymbol{\theta}}$ is calculated as $100^{-1} \sum_{d=1}^{100} \|\widehat{\boldsymbol{\theta}}_d - \boldsymbol{\theta}_{true}\|/p$, where $\boldsymbol{\theta}_{true} = (1, 1, 1, 1, 0.5, 0.5, -0.5, -0.5)/\sqrt{5}$ and $\boldsymbol{\gamma}_{true} = (1, 1, 1, 1, 0.5, 0.5, 0.5, 0.5)\ /\sqrt{5}$.

error for all parameters. The average loss in efficiency for the DPMM method in estimating the parameter $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ is 0.072 and 0.042 respectively. When the true distribution of error is scaled gamma, a skewed distribution, the average gain in efficiency of the DPMM method in estimating $\boldsymbol{\theta}$ is substantial and is approximately 1.325 times that of the normal method and 1.727 times that of the semiparametric kernel method. When the true distribution is a mixture of normals which represents a case when the true distribution can have potential outliers, DPMM attains almost

	Mean function($\boldsymbol{m}$)				Standard deviation($\boldsymbol{s}$)			
	LLK	Normal	DPMM	LLK2	LLK	Normal	DPMM	LLK2
$\epsilon \sim$ Normal								
$n = 200$	0.404	0.348	0.355	0.693	0.295	0.087	0.124	0.403
$n = 500$	0.261	0.288	0.293	0.699	0.111	0.084	0.104	0.533
$n = 1000$	0.223	0.243	0.249	0.227	0.163	0.089	0.099	0.371
Avg. Efficiency	**0.975**	**0.979**		**1.749**	**1.698**	**0.803**		**4.041**
$\epsilon \sim$ Scaled Gamma								
$n = 200$	0.319	0.342	0.332	0.325	0.113	0.094	0.093	0.739
$n = 500$	0.228	0.284	0.277	0.247	0.108	0.098	0.074	0.424
$n = 1000$	0.198	0.247	0.236	0.226	0.092	0.087	0.052	0.229
Avg. Efficiency	**0.874**	**1.034**		**0.944**	**1.481**	**1.336**		**6.937**
$\epsilon \sim$ Mixture of normals								
$n = 200$	0.287	0.361	0.353	0.358	0.223	0.087	0.080	0.934
$n = 500$	0.283	0.285	0.277	0.349	0.179	0.085	0.078	0.366
$n = 1000$	0.228	0.261	0.253	0.200	0.139	0.081	0.070	0.310
Avg. Efficiency	**0.912**	**1.028**		**1.022**	**2.356**	**1.176**		**6.932**

Table 3.2: The table shows the "root mean squared error" for mean function ($\boldsymbol{m}$) and standard deviation function ($\boldsymbol{s}$), which is evaluated for each of the 100 datasets under 3 different error distributions, namely, (1) Normal $(0, 1)$, (2) Scaled Gamma (mean = 0, sd = 1) and (3) Laplace (mean = 0, scale = 1). For each error distribution, we compare the performance of the new methods, (I) Normal and (II) DPMM with the local linear kernel method (LLK, when the starting value is closer to true value and LLK2, when the starting values are chosen by Principal Hessian Direction Method). RMSE for $\widehat{\boldsymbol{m}}$ is calculated as $100^{-1}\sum_{d=1}^{100}\|\widehat{\boldsymbol{m}}_d - \boldsymbol{m}_{true}\|$, where $\boldsymbol{m}_{true} = \sin\{\pi(X_i^{\mathrm{T}}\boldsymbol{\theta} - A)/(B-A)\}$ and $\boldsymbol{s}_{true} = \{0.2 + (X_i^{\mathrm{T}}\boldsymbol{\gamma})^2/8\}$, A and B are constants defined in section 3.3.1.

2 times more efficiency in estimating $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ than the other methods. Also the more robust method DPMM achieves substantial amount of efficiency with respect to the other methods in all the cases. With respect to the local linear kernel method, DPMM method gains almost 2.2 times root mean square efficiency for the mixture of normal case. The root mean square error of each method decreases as the sample size increases for all the parameters.

In Figure 3.1, we compare the modeling of all the methods for sample size equals

to 500. The estimation of the mean function by all the methods are almost unbiased as they are almost indistinguishable from the true model. The estimation of the standard deviation function is usually more difficult than the mean function if the error distribution is not correctly specified. P-spline fit using the DPMM method is closer to the true standard deviation function yielding significantly lesser amount of bias than the other methods. In Figure 3.2, we present the precision of the estimation of the distribution of the error for increasing sample size for different error models by the DPMM method.

3.4 Air Pollution Data

We use the NMMAPS (National Morbidity Mortality Air Pollution Study) database which contains daily mortality, weather and pollution data for 1987-2000. In the lower atmosphere (troposphere), ozone (O_3) is the most important photochemical oxidants. Controlled exposure studies on human and animals have provided evidence that ozone can cause adverse health effects both in short term and long term exposure. So in this study we want to model the association between the mean ozone level and levels of other 7 pollutants such as, *temperature, relative humidity, mean* CO_2 *level, mean* PM_{10} *level, mean* SO_2 *level, daily humidity range* and *daily temperature* for the year 1997. After eliminating one day with missing observations, we use observations from $n = 364$ days. All the levels of pollutants has been standardized with respect to their sample mean and sample standard deviation to do the modeling. The normal probability plot of deviance residuals show substantial departure from normality towards the right tail (Figure 3.3). Shapiro-Wilk test of normality on the residuals provides a pvalue of 0.0054 indicating a significant deviation from normality. So using a robust method (DPMM) could be more efficient than the normal method in this dataset. To do the local linear method we used the estimates

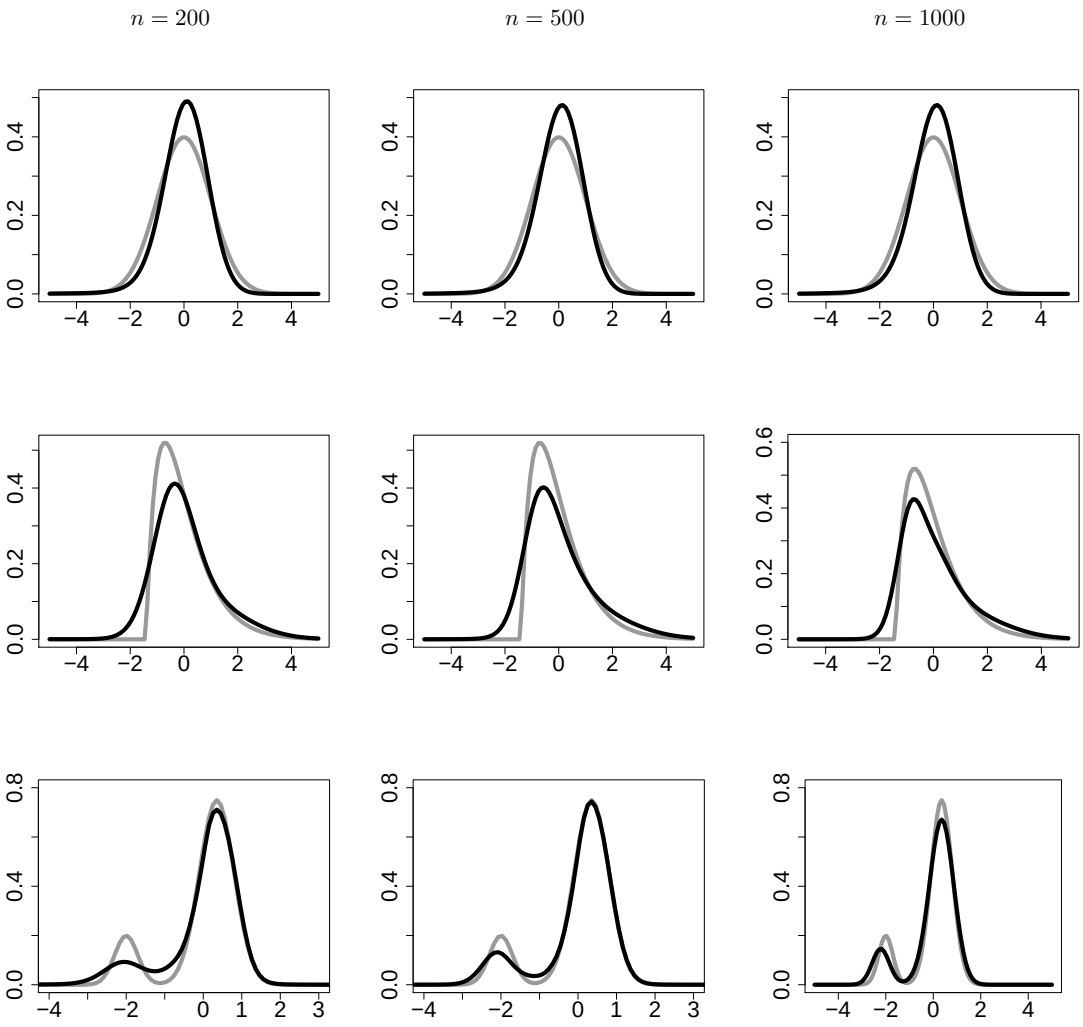

Figure 3.2: Estimation of the density of error by DPMM method for $n = 200$, 500, 1000. The first row shows the result when the true density is "Normal (0,1)", the second row corresponds to "scaled Gamma (mean = 0, sd = 1)" and the third row corresponds to "mixture of two normal distributions (mean = 0, sd = 1)". The "grey solid line" represents the true density, and the "black solid line" represents the DPMM estimation.

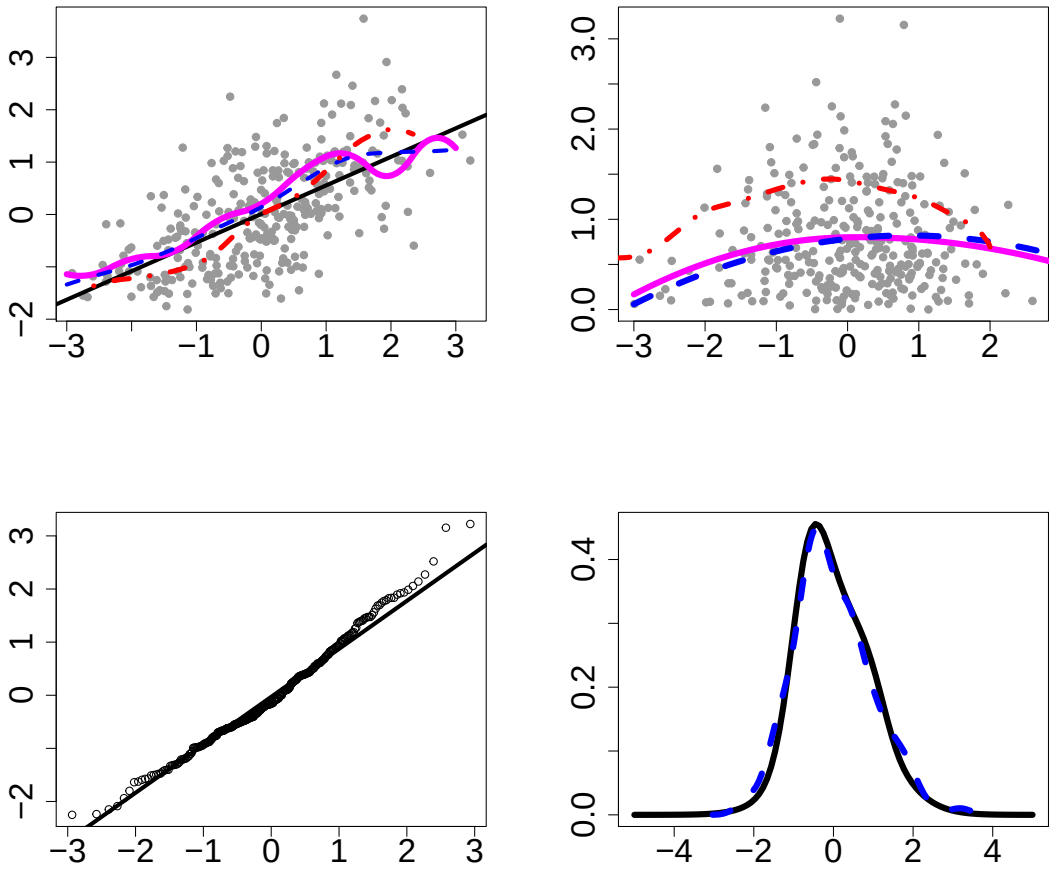

Figure 3.3: Summarizes the result of the Air Pollution data in Section 3.4. The first row shows the modeling of the mean level of ozone (left) and standard deviation of the ozone (right) with respect to the other pollutants. The "grey dots" in the mean function estimation are the true data points of mean ozone level. The "grey dots" in the variance function is the absolute residuals after mean modeling. The DPMM estimation is denoted by "magenta solid" line, the Normal method estimation is denoted by "blue dashed line", ordinary least square regression by "black solid line" and the local linear kernel method is denoted by "red dot-dashed" line. The second row represents the qqplot of the residuals (left) and the estimation of the density (right) of the residuals by DPMM method (black solid line) and kernel method (blue dashed line).

from the DPMM method as initial values for the iterations.

In Figure 3.3 we model the distribution of the error by DPMM method which looks asymmetric. We model the mean and the standard deviation function in Figure

3.3 and find a significant positive association between the mean ozone level and the levels of other pollutants. We also did an ordinary least square regression to model the mean ozone level and found that the estimated line is quite to the estimated curves from the semiparametric methods. In Table 3.3, we summarize the effects of the other pollutants by estimating $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$. Levels of the *mean temperature* has substantial positive effects on the mean ozone level. Whereas the mean CO_2 and *relative humidity* has significant negative effects. The nature of the results are fairly consistent with the findings by Ciaula and Bilancia (2015). A strong positive association between PM_{10} and the ozone level can be inferred from all the methods. We calculate the standard errors of the estimates by 100 Bootstrap samples where we found that the robust method DPMM obtains significant lower standard errors for the single-index parameters than the other methods.

3.5 Discussion

In this paper we attempt to model both the mean and variance function by single-index model, a popular multivariate semiparametric model with a powerful dimension reduction quality. To our best knowledge, one existing work by Lian, et al. (2014) addresses the issue by local linear kernel approach. In the numerical examples, we found that the method is quite sensitive to the starting values and thus can have serious issues in practical applications. We have an entirely different approach based on random effects B-spline methodology of mean and variance function, with flexible truncated Dirichlet process mixture of normals for the regression errors. We use Bayesian computation to fit, because the structure makes such computation simple and straightforward. In simulation, we found that the precision of the estimates largely increases under our method compared to that of the kernel approach. Even with initial values selected by simple methods like ordinary ridge regression

Parameters	Method	Mean Temp	Relative Humidity	Mean CO_2	Mean PM_{10}	Mean SO_2	Humidity Range	Temp Range
$\widehat{\boldsymbol{\theta}}$	**Normal**	**0.334**	**-0.251**	**-0.650**	**0.415**	**-0.335**	**-0.040**	**0.342**
	$\widehat{se}$	0.018	0.011	0.004	0.007	0.010	0.092	0.012
	DPMM	**0.226**	**-0.403**	**-0.533**	**0.460**	**-0.394**	**-0.066**	**0.362**
	$\widehat{se}$	0.009	0.013	0.018	0.009	0.017	0.084	0.020
	LLK2	**0.543**	**-0.176**	**-0.600**	**0.361**	**-0.264**	**0.090**	**0.325**
	$\widehat{se}$	0.013	0.024	0.006	0.013	0.016	0.041	0.015
$\widehat{\boldsymbol{\gamma}}$	**Normal**	**0.632**	**-0.458**	**-0.507**	**0.024**	**-0.318**	**0.161**	**-0.077**
	$\widehat{se}$	0.039	0.057	0.011	0.100	0.042	0.052	0.074
	DPMM	**0.242**	**-0.514**	**-0.326**	**0.331**	**-0.359**	**0.422**	**-0.392**
	$\widehat{se}$	0.015	0.015	0.052	0.037	0.033	0.036	0.044
	LLK2	**0.247**	**-0.584**	**-0.582**	**-0.052**	**-0.499**	**0.079**	**-0.002**
	$\widehat{se}$	0.051	0.007	0.008	0.063	0.011	0.044	0.106

Table 3.3: Summary of the estimates of single index parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ (in bold) denoting the effects of the other Air pollutants on the mean ozone level. The standard errors ($\widehat{se}$) are based on 100 bootstrap samples. We compare our methods based on our *Method I (Normal)* and *Method-II (DPMM)* models with that of the *local linear kernel method (LLK2)* of Lian, et al. (2014). Here Temperature is denoted as "Temp".

and principal hessian direction method (Li, 1991), our method yields consistent and efficient estimations in 5000 MCMC steps which takes about 15 mins in R using the Normal method and about 30 mins using the mixture of Normals. However, whether single-index model can be applicable to large scale problems where $n < p$ needs future study. An important extension of the current work and the subject of an ongoing research project is to find an integrated model which can effectively do variable selection of pertinent covariates to do single-index regression.

4. SINGLE-INDEX MODEL FOR SECONDARY ANALYSIS IN CASE-CONTROL STUDIES

4.1 The Model Framework

4.1.1 Background

Let the disease status be denoted as D with $D = 1$ denoting a case and $D = 0$ denoting a control. For $d = 0, 1$, let $\pi_d = \text{pr}(D = d)$, the probability that $D = d$ in the population. The disease status D is related to covariates $(Y, \boldsymbol{X})$ through the linear model

$$\text{pr}(D = d|\boldsymbol{X}, Y) = H(d, \boldsymbol{x}, y) = \frac{\exp\{d(\alpha_0 + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\alpha}_1 + y\alpha_2)\}}{1 + \exp\{\alpha_0 + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\alpha}_1 + y\alpha_2\}} \tag{4.1}$$

where $\boldsymbol{\alpha} = (\alpha_0, \boldsymbol{\alpha}_1, \alpha_2)$. In the secondary analysis, we seek to understand the regression relationship between covariate Y and multivariate covariate $\boldsymbol{X}$ in the true population based on n_1 number of cases and n_0 number of controls. We write $n = n_0 + n_1$ and introduce the parameter $\kappa = \alpha_0 + \log(n_1/n_0) - log(\pi_1/\pi_0)$. This reparametrization has the advantage that we can identify κ and $(\boldsymbol{\alpha}_1, \alpha_2)$ from a logistic analysis of D on (Y, X), although we cannot identify α_0 (Prentice and Pyke, 1979; Chatterjee and Carroll, 2005) from such logistic regression alone. We are interested in the following model

$$Y = m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}) + s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\epsilon \tag{4.2}$$

where, $m(\cdot)$ and $s(\cdot)$ are unkown flexible function based on unknown p-dimensional index vectors $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, The error ϵ are independent and identically distributed ac-

cording to some density $f_\epsilon(\cdot)$, with only restrictions $E(\epsilon|\boldsymbol{X}) = 0$ and $E(\epsilon^2|\boldsymbol{X}) = 1$ to ensure identifiability. Hence, the conditional heteroscedasticity is only explained through the variance function $s^2(\cdot)$. To avoid identifiability issues, we need additional restrictions on the single-index vectors, specifically, $\|\boldsymbol{\theta}\| = \|\boldsymbol{\gamma}\| = 1$.

4.1.2 Modeling the Mean Function $m(\cdot)$

Many of the problems in case-control studies involve nonlinear relationship between covariates that are difficult to model parametrically. There is a clear imperative to be able to handle such nonlinear relationships effectively through more flexible techniques. Although there are several methods for constructing "smooth" function, in this paper we focus on *penalized splines*. It has the attractiveness of being a relatively straightforward extension of linear mixed model (O' Sullivan, 1986; Eilers and Marx, 1996).

Given $\boldsymbol{\theta}$, a general approach discussed by Eilers and Marx (1996) is to model the regression function $m(\cdot)$ by a linear combination of *basis splines* or *B-splines* with fixed knots and smoothing parameter ρ_1, better known as *penalized splines* or *P-splines*. Let $\mathcal{B}_1(t) = \{\mathcal{B}_{1j}(t)\}_{j=1}^{M}$, where $M = K + q - 2$ be a q^{th} order B-spline with K fixed knots. The coefficient of the k^{th} knot is denoted by β_j for $j = 1, ..., M$. The mean function is evaluated at the "transformed design points" $t_i = \boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\theta}}$ for $i = 1, ..., n$ as

$$\widehat{m}(t_i) = \sum_{j=1}^{M} \mathcal{B}_{1j}(t_i)\beta_j. \tag{4.3}$$

Let D be a fixed, symmetric and a positive semi-definite N-dimensional matrix. The

penalized least square estimator $\widehat{\boldsymbol{\beta}}(\rho_1, \boldsymbol{\theta})$ for the mean function minimizes

$$\sum_{i=1}^{n} \{Y_i - \mathcal{B}_1(t_i)^{\mathrm{T}} \boldsymbol{\beta}\}^2 + \rho_1 \boldsymbol{\beta}^{\mathrm{T}} D \boldsymbol{\beta},$$

Here $\boldsymbol{\beta} = \{\beta_1, \beta_2, ..., \beta_M\}^{\mathrm{T}}$ and $D = D_r^{\mathrm{T}} D_r$ is the matrix representation of the difference operator Δ_r of order r defined by Eilers and Marx (1996). For simplicity of notation, we define $\widetilde{\mathbf{X}}_{\boldsymbol{\theta}} = \mathcal{B}_1(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{\theta})$ to obtain the least square estimators for $\boldsymbol{\beta}$

$$\widehat{\boldsymbol{\beta}}(\rho_1, \boldsymbol{\theta})_{LS} = (\widetilde{\mathbf{X}}_{\boldsymbol{\theta}}^{\mathrm{T}} \widetilde{\mathbf{X}}_{\boldsymbol{\theta}} + \rho_1 D)^{-1} \widetilde{\mathbf{X}}_{\boldsymbol{\theta}}^{\mathrm{T}} Y.$$

We can use $\widehat{\boldsymbol{\beta}}_{LS}$ as a potential starting value for the MCMC iteration.

4.1.3 Modeling the standard deviation function $s(\cdot)$

In many cases, the assumption of constant conditional variance is unrealistic. There are instances in biology where the effect of a treatment is to cause an increase in variance rather than an increase in mean. A comprehensive review of heteroscedasticity is given in Carroll and Ruppert (1988). There are instances that even with various transformation, it is hard to stabilize the variance. We take up the flexible approach based on penalized splines to model the variance function as another single-index model with respect to $\boldsymbol{X}$.

In the Bayesian approach, estimation of a flexible variance function $s(\cdot)$ and the corresponding single-index vector $\boldsymbol{\gamma}$ is straightforward. We only need to specify the prior distribution of the parameters. Although there are many ways to implement variance function, one such example of modeling log-transformed variance function by flexible mixture of splines are abundant in the literature, for example Yau and Kohn (2003), Liu, et al. (2006). In this article, for fixed $\boldsymbol{\gamma}$, we model the variance function by a positive mixture of B-splines of order q with K_1 knots at transformed

points where we use the exponential function to ensure that the variance function is positive $w_i = \boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}$ as

$$s(w_i) = \sum_{j=1}^{M_1} \mathcal{B}_{2j}(w_i)\exp(\xi_j) \tag{4.4}$$

where, $M_1 = K_1+q-2$. For simplicity, we define $\widetilde{\mathbf{X}}_{\boldsymbol{\gamma}} = \mathcal{B}_2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})$. Then a reasonable starting value for the P-spline coefficient $\boldsymbol{\xi}$ of the variance function is

$$\exp\left\{\widehat{\boldsymbol{\xi}}(\rho_2,\boldsymbol{\gamma})_{LS}\right\} = |(\widetilde{\mathbf{X}}_{\boldsymbol{\gamma}}^{\mathrm{T}}\widetilde{\mathbf{X}}_{\boldsymbol{\gamma}} + \rho_2 D)^{-1}\widetilde{\mathbf{X}}_{\boldsymbol{\gamma}}^{\mathrm{T}}\widehat{r}^2|,$$

4.1.4 Case-control Likelihood

The conditional distribution of Y given $\boldsymbol{X}$ is modeled as $f_\epsilon\{y-m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}), s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\}$. For the case-control studies, Jiang *et al.* (2006), Chen *et al.* (2008) and Lin and Zeng (2009) derived the efficient profile likelihood. Write $(\Omega = \kappa, \boldsymbol{\alpha}_1, \alpha_2)$. The joint density of D, Y given $\boldsymbol{X}$ is

$$g(d,y|\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\gamma},\Omega) = f_\epsilon[\{y-m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})\}/s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})]\frac{\exp\{d(\alpha_0+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\alpha}_1+y\alpha_2)\}}{1+\exp(\alpha_0+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\alpha}_1+y\alpha_2)}$$

The semiparametric efficient retrospective profile likelihood for $Y|\boldsymbol{X}$ when the distribution of Y given $\boldsymbol{X}$ is specified is

$$\mathcal{L}(Y|\boldsymbol{X}, D=d, \boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\alpha}) = \frac{g(d,y|\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\alpha})}{\int g(d,t|\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\alpha})dt} \tag{4.5}$$

Under homoscedastic errors, Wei *et al.* (2013) showed that the scores function of the regression parameters yield 0 under the true value only if the density f_ϵ is specified properly. So this motivates our search for a robust nonparametric estimation method

under high-dimensional set up. The notation Normal$(\cdot|\mu, \sigma)$ is used to denote a normal distribution with mean μ and standard deviation σ. We denote $\phi(\cdot)$ as the standard normal distribution. In this article, we model f_ϵ by two ways

I. Errors are normally distributed with mean 0 and standard deviation 1.

$$f_\epsilon(\cdot) = \phi(\cdot).$$

II. A more robust model based on finite mixture of Normals with mean 0

$$f_\epsilon(\cdot) = \sum_{k=1}^{C} w_k\{p_k N(\cdot|\mu_{k1}, \sigma_{k1}) + (1 - p_k)N(\cdot|\mu_{k2}, \sigma_{k2})\},$$

where $\sum_{k=1}^{C} w_k = 1$ and $p_k\mu_{k1} + (1 - p_k)\mu_{k2} = 0$ for $k = 1, \cdots, C$.

4.2 Identifiability Issues and Rare disease approximation

It is typical in case-contol studies that the disease rate, $\pi_1 = \text{pr}(D = 1)$ is not known. Also under case-control set up, the logistic intercept α_0 is not identifiable (Prentice and Pyke, 1979). So under frequentist set up, the score function of α_0 produces extremely unstable results, Wei *et al.* (2013). That motivates us to substitute the complete likelihood by an approximate likelihood which can yield almost consistent results without estimating α_0.

Since case-control studies are almost inevitably conducted for rare outcomes, the rare disease approximation is natural in most applications. Although the word "rare" has no specific definition but it is certainly 1% or less. In other words, a disease is *rare* if $\text{pr}(D = 1) = \pi_1 \approx 0$. In the literature most researchers like Piegorsch *et al.* (1994), Epstien and Satten (2003), Lin and Zeng (2006), Modan *et al.* (2001), Zhao *et al.* (2003), Kwee *et al.* (2007), Lin and Zeng (2009), Hu *et al.* (2010), Wei *et al.*

(2013) and Rahman *et al.* (2014) used the following approximation to find almost consistent results.

$$\text{pr}(D=1|Y,\boldsymbol{X}) = \frac{\exp(\alpha_0 + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\alpha}_1 + y\alpha_2)}{1+\exp(\alpha_0 + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\alpha}_1 + y\alpha_2)} \approx \exp(\alpha_0 + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\alpha}_1 + y\alpha_2) \qquad (4.6)$$

Equivalently, under rare disease, $\text{pr}(D=0|Y,\boldsymbol{X})$ becomes approximately equal to 1.

4.2.1 Standard Method: Using Only Controls

Due to overrepresentation of cases in the sample, case-control samples are not random sample from the true population. So application of standard method of estimation leads to biased results. Case-controls are typically used to study rare disease. Otherwise, with disease rate π_1 being unknown the nonparametric estimation is not identifiable. So in this case, invoking rare disease approximation has been a popular way to solve the problem (Nagerkele, *et al.*, 1995; Jiang, *et al.* 2006; Lin and Zeng, 2010; Rahman, 2015). Under rare disease approximation, when $\pi_1 \approx 0$ or when $\pi_1 \approx 1$, the population of control can be regarded approximately equal to the true population. Then the standard methods based on i.i.d samples applied to only controls lead to approximately consistent result.

When error distribution is $f_\epsilon(\cdot) = \text{Normal}(\cdot|0,1)$, the likelihood based only on controls is

$$L_{con,0,1}(y|d,\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\gamma}) = \phi[\{y - m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})\}/s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})]^{1-d}$$

So taking the logarithms and summing over the observed data, the log-likelihood function is

$$\mathcal{L}_{con,0,1}(y|d,\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\gamma}) = \sum_{i=1}^{n}(1-d_i)\log\Big(\phi[\{y_i - m({\boldsymbol{X}_i}^{\mathrm{T}}\boldsymbol{\theta})\}/s({\boldsymbol{X}_i}^{\mathrm{T}}\boldsymbol{\gamma})]\Big) \qquad (4.7)$$

When the error distribution is a finite mixture model of Normals,

$$\mathcal{L}_{con,mixture}(y|d,\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\gamma}) = \sum_{i=1}^{n}(1-d_i)\log[\sum_{k=1}^{C} w_k\{p_k N(\cdot|\mu_{k1},\sigma_{k1}) + (1-p_k)N(\cdot|\mu_{k2},\sigma_{k2})\}] \quad (4.8)$$

4.2.2 Approximate Efficient Likelihood Using Normal Errors

Define $g_{approx}(y|d,\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\gamma},\Omega) = f_\epsilon[\{y - m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})/s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})]\exp\{d(\kappa + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\alpha}_1 + y\alpha_2)\}$. Under the rare disease approximation the working likelihood can be written as

$$L_{approx}(Y|\boldsymbol{X}, D=d,\boldsymbol{\theta},\boldsymbol{\gamma},\Omega) = \frac{g_{approx}(y|d,\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\gamma},\Omega)}{\int g_{approx}(t|d,\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\gamma},\Omega)dt}$$

When $f_\epsilon(\cdot)$ is Normal$(0,1)$, the approximate case-control likelihood is

$$L_{norm}(y|d,\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\gamma},\alpha_2) = \exp[d\alpha_2 y - d\{\alpha_2 m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}) + \alpha_2^2 s^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})/2\}]\phi\{\frac{y - m(X^{\mathrm{T}}\boldsymbol{\theta})}{s(X^{\mathrm{T}}\boldsymbol{\gamma})}\} \quad (4.9)$$

So, the rare disease approximation in equation (4.6) allows us to avoid the identifiability issue of logistic intercept. And it also enables us to utilize the entire case-control sample in the likelihood.

4.2.3 Approximate Efficient and Robust Likelihood Using Finite Mixture of Normals

Recall that in section 4.1.4, we discussed that the score under case-control likelihood donot yield 0 under true parameter if the conditional distribution of $f(Y|\boldsymbol{X})$ is misspecified. So our goal of this paper is to find a robust and efficient estimation of $m(\cdot)$ and $s(\cdot)$ using the entire case-control data. Rahman *et al.* (2015) showed that finite mixture of Normal can be suitably used to capture any considerable deviation from normality while estimating mean and variance function for single-index model.

In particular, they showed that for heavy tail distribution, finite mixture of Normals can achieve a substantial gain in efficiency under i.i.d. set up.

Write $\boldsymbol{\Psi} = (\kappa, \boldsymbol{\alpha}_x, \alpha_y, \boldsymbol{\theta}, \boldsymbol{\gamma}, \{\mu_{k1}\}_{k=1}^C, \{\sigma_{k1}\}_{k=1}^C, \{\sigma_{k2}\}_{k=1}^C)$. Define $\mathcal{Q}_1(d, \alpha_2, \boldsymbol{\Psi}) = \sum_{k=1}^C w_k p_k \exp\{d\alpha_2 s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\mu_{k1} + d\alpha_2^2 s^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma}) s_{k1}^2/2\}$ and $\mathcal{Q}_2(d, \alpha_2, \boldsymbol{\Psi}) = \sum_{k=1}^C w_k(1-p_k) \exp\{d\alpha_2 s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\mu_{k2} + d\alpha_2^2 s^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma}) s_{k2}^2/2\}$. The approximate efficient case-control likelihood under finite mixture of Normal is

$$\begin{aligned} L_{robust}(y|d, \boldsymbol{X}, \boldsymbol{\Psi}, \alpha_2) &= \frac{\exp\{d\alpha_2 y - d\alpha_2 m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})\}}{\mathcal{Q}_1(d, \alpha_2, \boldsymbol{\Psi}) + \mathcal{Q}_2(d, \alpha_2, \boldsymbol{\Psi})} \Big[\times \sum_{k=1}^{C} w_k \{p_k N(\{y|\mu_{k1}, \sigma_{k1}) \\ &\quad + (1-p_k) N(y|\mu_{k2}, \sigma_{k2})\}\Big] \end{aligned} \tag{4.10}$$

4.3 Estimation Methods

4.3.1 Using Only Controls

The controls can be regarded as random sample of the true population if the disease is rare. Therefore the true regression function can be estimated using only the controls using the likelihood in (4.7) and (4.8). Hence the method is similar to i.i.d samples developed by Rahman *et al.* (2016). The method is as follows.

(a) Starting values for index vectors $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are obtained by applying Principal Hessian Direction matrix (Li, 1992) on the controls.

(b) Design the non-informative prior for all parameters.

(c) Update all the parameters by MCMC based only on controls.

(d) Obtain the posterior estimates $\widehat{\boldsymbol{\beta}}_{con}$, $\widehat{\boldsymbol{\xi}}_{con}$, $\widehat{\boldsymbol{\theta}}_{con}$ and $\widehat{\boldsymbol{\gamma}}_{con}$.

4.3.2 Efficient Estimation Using Entire Case Control Data

Rahman *et al.* (2015) showed that Bayesian MCMC procedure can attain a practical, consistent and efficient estimation of mean and variance single-index model for

i.i.d samples. The method is successful in obtaining the consistent estimates of the true mean and variance function without depending heavily on the initial values unlike other frequentist method. That motivates us to develop a Bayesian methodology based on finite mixture error model. The lay-out of the efficient estimation strategy under case-control set up is as follows.

(a) Estimate the true logistic regression parameters κ, $\boldsymbol{\alpha}_1$ and α_2 by ordinary logistic regression of D on $(Y, \boldsymbol{X})$. This is a good starting value for because the ordinary logistic regression in a case-control study consistently estimates $(\boldsymbol{\alpha}_1, \alpha_2)$ and κ (Prentice and Pyke, 1979; Chatterjee and Carroll, 2005). We denote the estimates as $\widehat{\kappa}_{init}$ and $(\widehat{\boldsymbol{\alpha}}_{1,init}, \widehat{\alpha}_{2,init})$.

(b) Use $\widehat{\boldsymbol{\theta}}_{con}, \widehat{\boldsymbol{\gamma}}_{con}$ as the starting values for the index parameter $\boldsymbol{\theta}, \boldsymbol{\gamma}$.

(c) Use $\widehat{\boldsymbol{\beta}}_{con}, \widehat{\boldsymbol{\xi}}_{con}$ as the starting values for the regression parameter $\boldsymbol{\beta}, \boldsymbol{\xi}$.

(d) Use the same non-informative prior for each of the parameters.

(e) Since full posterior conditional distribution of the parameters are not straightforward, update all the parameters by MCMC algorithm.

4.4 Prior Specification

In this section we develop the fully Bayesian structure of the model by placing priors on all the parameters. In this paper we use the definition of the inverse-gamma and gamma distribution from Berger (1985),

$$\mathrm{IG}(A, B) = \frac{1}{\Gamma(A) B^A x^{A+1}} \exp\Big\{ -1/(Bx) \Big\} I_{(0,\infty)}(x)$$

and

$$\text{Gamma}(A, B) = \frac{1}{\Gamma(A)B^A} x^{A-1} \exp\Big\{ -x/B \Big\} I_{(0,\infty)}(x).$$

The generic notation p_0 is used for specifying priors for parameters and hyperparameters.

1. **Single-index Vectors:** Without loss of generality, we fix the first component of $\boldsymbol{\theta}$ to 1, while rest of the component of $\boldsymbol{\theta}$ as $\boldsymbol{\theta}_{-1} = (\theta_2, ..., \theta_p)$. We use the eigenvector corresponding to the maximum eigenvalue of $\widehat{\Lambda}$ as a *starting value* for $\boldsymbol{\theta}_{-1}$. Similarly we denote $\boldsymbol{\gamma}_{-1} = (\gamma_2, ..., \gamma_p)$ and calculate the *starting value* for $\boldsymbol{\gamma}_{-1}$ using the maximum eigen value of $\widehat{\Gamma}$. For each MCMC iteration step, we fix $\theta_1 = \gamma_1 = 1$ and specify normal prior on the rest of the components of the vectors

$$p_0(\boldsymbol{\theta}_{-1}) = \text{Normal}(\cdot|\boldsymbol{\theta}_{\text{prior}}, \Sigma_{\boldsymbol{\theta}}),$$

$$p_0(\boldsymbol{\gamma}_{-1}) = \text{Normal}(\cdot|\boldsymbol{\gamma}_{\text{prior}}, \Sigma_{\boldsymbol{\gamma}}),$$

   where $\boldsymbol{\theta}_{\text{prior}}, \boldsymbol{\gamma}_{\text{prior}}$, $\Sigma_{\boldsymbol{\theta}}$ and $\Sigma_{\boldsymbol{\gamma}}$ are pre-specified constants.

2. **Mean function:** Let $\boldsymbol{\beta} = \{\beta_1, \beta_2, ..., \beta_M\}^{\text{T}}$ and $D = D_r^{\text{T}} D_r$ is the matrix representation of the difference operator Δ_r of order r defined by Eilers and Marx (1996). We model the mean function by a B-spline with smoothness

inducing prior on the coefficients, given as

$$\begin{aligned} p_0(\boldsymbol{\beta}) &\propto \rho_1^{M/2} \exp\{-\rho_1 \boldsymbol{\beta}^{\mathrm{T}} D \boldsymbol{\beta}/2\}, \\ \rho_1 &\sim \text{Gamma}(A_{\rho 1}, B_{\rho 1}). \end{aligned}$$

The prior $p_0(\boldsymbol{\beta})$ induces smoothness in the coefficients because it penalizes $\sum_{j=r+1}^{n} (\Delta_r \beta_j)^2 = \boldsymbol{\beta}^{\mathrm{T}} D \boldsymbol{\beta}$, the sum of squares of the second order difference in $\boldsymbol{\beta}$.

3. **Variance function:** A flexible Bayesian model for the variance function with smoothness inducing priors on the coefficients is

$$\begin{aligned} p_0(\boldsymbol{\xi}) &\propto \rho_2^{M_1/2} \exp\{-\rho_2 \exp(\boldsymbol{\xi})^{\mathrm{T}} D \exp(\boldsymbol{\xi})/2\}, \\ \rho_2 &\sim \text{Gamma}(A_{\rho 2}, B_{\rho 2}). \end{aligned}$$

Here ρ_2 is the smoothing parameter for the variance function. Larger values of ρ_2 with smaller number of knots K_1 imposes stronger penalty resulting in smoother variance function.

4. **Finite mixture of Normal model parameters:** For modeling a density $f(\cdot)$, a DPMM with concentration parameter α, base measure P_0 and mixture components coming from a parametric family $\{f_c(\cdot|\phi) : \phi \sim P_0\}$, can be specified as

$$f(\cdot) = \sum_{k=1}^{\infty} \pi_k f_c(\cdot|\phi_k), \quad \phi_k \sim P_0, \quad \pi_k = \pi_k^* \prod_{i=1}^{k-1} (1 - \pi_j^*), \quad \pi_k^* \sim \text{Beta}(1, \alpha),$$

In the literature, this construction of random mixture weights $\{\pi_k\}_{k=1}^{\infty}$, illustrated first by Sethuraman (1994), is known as Stick-breaking representation,

hence $\pi \sim \text{Stick}(\alpha)$. Lo (1984) showed that a DPMM of normal density, that is $f_c(\cdot) = \text{Normal}(\cdot)$ is dense in the space of densities with respect to Lebesgue measure. Hence, DPMMs of normals are popular for modeling densities (Escobar and West, 1995; West, et al. 1994). In the context of regression analysis, moment constraint infinite mixture models of Normal has been considered by Griffin and Steel (2010) and Pelenis (2014) so that the mean of the error can be restricted at 0. Drawing inspiration from them, we let $f_c(\cdot)$ to be a two component mixture of normal as

$$f_c(\cdot|p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = p\text{Normal}(\cdot|\mu_1, \sigma_1^2) + (1-p)\text{Normal}(\cdot|\mu_2, \sigma_2^2),$$

subject to the moment constraint

$$p\mu_1 + (1-p)\mu_2 = 0. \tag{4.11}$$

We define $\mu^* = -p\mu_1/(1-p)$ and impose the constraint (4.11) to obtain the complete DPMM prior for the density function of ϵ

$$\begin{aligned} f_\epsilon(\epsilon) &= \sum_{k=1}^{\infty} \pi_k f_c(\epsilon|p_k, \mu_{k1}, \mu_k^*, \sigma_{k1}^2, \sigma_{k2}^2), \qquad (4.12) \\ \boldsymbol{\pi} &\sim \text{Stick}(\alpha), \qquad p_k \sim \text{Unif}(0,1), \\ \mu_k &\sim \text{Normal}(a_0, b_0), \qquad \sigma_{k1}^2, \sigma_{k2}^2 \sim \text{IG}(c_0, d_0). \end{aligned}$$

Here α, b_0, c_0 and d_0 are positive preassigned constants. We used blocked Gibb's sampler to sample the parameters from their posterior (Ishwaran and James, 2001).

DPMMs are essentially mixture models with a potential infinite number of

mixtures or 'clusters'. However, for a given data set of finite size, the number of active clusters exhibited by the data is finite. So instead of infinite mixture of normals, we can consider a finite mixture of densities upto a number C

$$f(\cdot) = \sum_{k=1}^{C} \pi_k f_c(\cdot|\phi_k), \quad \phi_k \sim P_0, \quad \pi_k = \pi_k^* \prod_{i=1}^{k-1}(1-\pi_j^*), \quad \pi_k^* \sim \text{Beta}(1,\alpha),$$

Screening techniques by Berkhof et al. (2003) and Li and Chen (2010) for number of components for the finite mixture model can be used to determine C.

4.5 Simulation

We performed a simulation studies both at and away the Gaussian model to assess the first nonparametric single-index model attempt in case-control studies. Our simulations indicate that our MCMC posterior estimates has small bias. We also show that our method achieves significant gain in efficiency when compared with using only controls while the approach that uses all the data but ignores the case-control sampling design suffers from bias. The result presented here are based on P-splines with 10 knots for mean function estimation and 5 knots for variance function estimation. In each case, 500 cases and 500 controls are generated with X_i generated from Unif$(0,1)$ and Y is simulated from the following model used by Carroll et al. (1997),

$$y_i = \sin\left\{\ \pi(X_i^{\mathrm{T}}\boldsymbol{\theta} - A)\right\} + \left\{0.2 + (X_i^{\mathrm{T}}\boldsymbol{\gamma})^2/8\right\}\epsilon_i, \tag{4.13}$$

where $A = \sqrt{3}/2 - 1.645/\sqrt{12}$.

We take $\boldsymbol{\theta}^{\mathrm{T}} = (1,1,1)/\sqrt{3}$ and $\boldsymbol{\gamma}^{\mathrm{T}} = (1,1,1)/\sqrt{3}$. We use two distribution for the distribution of errors ϵ_i, (I) Normal$(\epsilon_i|0,1)$ and (II) Laplace distribution with

mean 0 and standard deviation 1. The logistic regression model is $\text{pr}(D = 1|Y, X) = H(\alpha_0 + \alpha_y Y + \boldsymbol{\alpha}_x X)$ with $\alpha_y = 0.25$ and $\boldsymbol{\alpha}_x{}^{\text{T}} = (1, 1, 1)$. We use $\alpha_0 = -5.2$ to simulate from the true population with disease rate approximately 3%. We contrasted four methods, (a) the single-index model using entire data sets ignoring the case-control set up which is expected to be biased (ALL), (b) using only controls based on Normal density(CONT), (c) using the adjusted Likelihood based on Normal density (ANL) and (d) using the adjusted Likelihood based on mixture of normals (AMNL). To compare the performance of the methods, we simulated 100 datasets where we computed the root mean squared error (RMSE) and standard error (s.e) of the index vectors $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, mean function $m(\cdot)$ and standard deviation function $s(\cdot)$. For B simulated datasets, we computed the RMSE($\widehat{\boldsymbol{\theta}}_i$) of parameter $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_3)$ as $\sqrt{1/B \sum_{r=1}^{B} (\widehat{\theta}_{ir} - \theta_i)^2}$. For the mean and standard deviation function we calculate the root mean squared error in a predefined grid on $(0, 1)$ divided into 100 equal intervals. For a point z_i on the grid, we compute the RMSE of $\widehat{m}(z_i)$ as $\sqrt{1/B \sum_{r=1}^{B} \{\widehat{m}(z_i)_r - m(z_i)\}^2}$.

In Table 4.1 and Table 4.2, the "MSE Efficiency" of the new methods ANL, AMNL and ALL is calculated with respect to the control only method (CON) by (MSE Efficiency of the method)/(MSE Efficiency of the controls). For $\alpha_y = 0.00$ then Li, et al. (2010) observed that when Y is independent of D given $\boldsymbol{X}$, then the association between the covariates in the cases remains same as that for the underlying true population. So we present the result for $\alpha_y = 0.25$. Here the approach that uses all the data (ALL) as i.i.d. is biased in estimating the mean and standard deviation function. The estimates of $\boldsymbol{\theta}$ under "ALL" method is biased, see Table 4.1. Under true Normal model, our Bayesian method using adjusted Normal likelihood (ANL) is more efficient in estimating both $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$. The adjusted method is almost 1.5 times as efficient as that of using standard method on controls. In

	Normal model				Laplace Model			
	CONT	ANL	AMNL	ALL	CONT	ANL	AMNL	ALL
$\theta_1 = 0.577$								
Mean	0.581	0.580	0.586	0.768	0.574	0.572	0.576	0.774
s.e	0.007	0.008	0.009	0.004	0.019	0.012	0.018	0.004
MSE Eff		1.587	0.991	0.171		0.916	1.456	0.098
$\theta_2 = 0.577$								
Mean	0.567	0.571	0.578	0.465	0.599	0.595	0.585	0.464
s.e	0.027	0.019	0.024	0.016	0.015	0.016	0.007	0.016
MSE Eff		1.301	1.218	0.261		1.116	2.540	0.232
$\theta_3 = 0.577$								
Mean	0.573	0.576	0.577	0.440	0.558	0.575	0.577	0.431
s.e	0.028	0.023	0.028	0.018	0.004	0.005	0.001	0.018
MSE Eff		1.157	1.158	0.478		2.235	3.481	0.137
Mean function Efficiency								
		2.325	2.083	1.004		2.011	3.015	0.865

Table 4.1: Result of the simulation study for the single index parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ with $n_1 = 500$ cases and $n_0 = 500$ controls, and a disease rate of approximately 3%. To obtain the response, we consider two distributions, "Normal Model ($\epsilon \sim$ N(0, 1))" and "Laplace model ($\epsilon \sim$ Laplace(0,1)"). For 100 simulated data sets, we computed the mean of the estimates("Mean"), it's standard error ("s.e"), lower ("Lower) and upper ("Upper") 95% confidence intervals and the root-mean-squared error efficiency ("MSE Eff") compared with using only the controls. Our methods (ANL) and (AMNL) are contrasted with using (a) only controls with normal likelihood ("CONT") and (b) Entire case-control data with normal likelihood ("ALL").

estimating the mean function, both the adjusted likelihood based on Normal (ANL) and mixture of Normals (AMNL) are almost 2 times more efficient than that of using only controls (CON). However, while estimating the standard deviation function, a more difficult problem, our method based on mixture of normal is less biased and more efficient than other methods. When the error is away from Gaussian, for example, we consider the case when the error distribution is Laplace, we found that our method using normal likelihood lacks efficiency. Our adjusted method using finite mixture of normals is almost 1.5 times as efficienct than the adjusted normal

	Normal model				Laplace Model			
	CONT	ANL	AMNL	ALL	CONT	ANL	AMNL	ALL
$\gamma_1 = 0.577$								
Mean	0.680	0.628	0.677	0.619	0.533	0.612	0.518	0.488
s.e	0.014	0.009	0.022	0.015	0.071	0.062	0.041	0.089
MSE Eff		2.004	1.021	1.965		1.768	1.139	0.791
$\gamma_2 = 0.577$								
Mean	0.526	0.532	0.528	0.528	0.606	0.532	0.587	0.284
s.e	0.035	0.039	0.033	0.061	0.074	0.069	0.066	0.441
MSE Eff		0.970	0.977	0.661		1.970	2.878	0.168
$\gamma_3 = 0.577$								
Mean	0.510	0.565	0.511	0.576	0.570	0.681	0.620	0.748
s.e	0.012	0.047	0.005	1.715	0.125	0.117		0.211
MSE Eff		1.397	1.029	0.429		1.202	2.797	0.591
S.D. function Efficiency								
		1.152	1.809	0.473		1.182	1.462	0.846

Table 4.2: Result of the simulation study for the single index parameter $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$ with $n_1 = 500$ cases and $n_0 = 500$ controls, and a disease rate of approximately 3%. To obtain the response, we consider two distributions, "Normal Model ($\epsilon \sim$ N$(0, 1)$)" and "Laplace model ($\epsilon \sim$ Laplace(0,1)". For 100 simulated data sets, we computed the mean of the estimates ("Mean"), it's standard error ("s.e"), lower ("Lower) and upper ("Upper") 95% confidence intervals and the root-mean-squared error efficiency ("MSE Eff") compared with using only the controls. Our methods (ANL) and (AMNL) are contrasted with using (a) only controls with normal likelihood ("CONT") and (b) Entire case-control data with normal likelihood ("ALL").

likelihood method in estimating the mean function and almost 3 times efficient than that using only controls. In estimating the variance function, the adjusted method based on finite mixture of normal attains maximum efficiency with respect to using only controls and the normal likelihood method.

4.6 Secondary Analysis on NIH-AARP Data Diet and Health Study

We use the case-control data from NIH-AARP (National Institute of Health-American Association of Retired Persons to study on the association between the

BMI(Body Mass Index) and eating habit (like, Alcohol intake). This study was done in 1995-1996 to address the epidemiologic investigations of diet and cancer (Schatzkin, et al. 2001). The study is based on the responses from the baseline questionnaire to both men and women of age 50 years and over, an age when cancer occurrence is becoming more frequent. In this paper we focus on the breast cancer incidences of $n1 = 1025$ number of cases and $n0 = 1086$ of controls. "BMI" is recorded in kg/m^2, "Alcohol intake" is recorded in gm/day, "Age" in years and "Fat density" in percentages. In this data set, there was no missing data and we standardized each variables before using our methods.

Before applying our method, we do a preliminary analysis by fiting a quadratic model to model "BMI" by "Fat density", "Age" and "Alcohol intake". The pvalue for the quadratic part is $< 10^{-3}$. We compare 3 methods, (a) Using only controls(CON), (b) Using adjusted likelihood based on Normal errors(ANL), (c) Using adjusted likelihood based on Mixture of Normals(ANML).

The results are given in Table 4.3. We see in Table 4.3 that estimation of $\boldsymbol{\theta}$ for Age and Fat density is almost same for all the three methods. To compute the efficiency of the methods we compare the standard errors of the estimates. Efficiency of the robust method (AMNL) is higher than the rest of the approaches. The efficiency gain is almost 2 times that of the controls (CONTROLS) in estimating the mean function parameter $\boldsymbol{\theta}$(Table 4.3). While estimating the standard deviation function, the efficiency gain of Normal approach (ANL) is greater than that of the AMNL. For estimation of ALCOHOL effect, the standard errors of all the methods are quite high signifying a non-significant effect of Alcohol in modeling BMI. The pvalues for linear and quadratic model reveals very insignificant effect of Alcohol on BMI in this dataset. The effect of FAT DENSITY is quite high in all the methods, which matches with the preliminary linear model done of BMI and other covariates.

	$\boldsymbol{\theta}$			$\boldsymbol{\gamma}$		
	CONT	ANL	AMNL	CONT	ANL	AMNL
AGE						
Mean	-0.414	-0.406	-0.431	-0.461	-0.665	-0.652
s.e	0.071	0.092	0.073	0.198	0.191	0.138
Efficiency		0.771	0.973		1.037	1.435
ALCOHOL						
Mean	-0.221	0.198	0.014	-0.217	0.332	-0.050
s.e	0.301	0.172	0.145	0.243	0.109	0.306
Efficiency		1.742	2.076		2.229	0.794
FAT DENSITY						
Mean	0.883	0.892	0.902	0.861	0.669	0.756
s.e	0.129	0.051	0.049	0.195	0.085	0.136
Efficiency		2.519	2.584		2.294	1.434

Table 4.3: Result of NIH-AARP study when BMI is modeled by single-index function of "Age", "Alcohol Intake" and "Fat Density" from the 1000 cases of Breast Cancer and 1000 number of controls. For 100 simulated data sets, we computed the mean of the estimates("Mean"), it's standard error ("s.e"), lower ("Lower) and upper ("Upper") 95% confidence intervals and the standard error efficiency ("Efficiency") compared with using only the controls. Our methods (ANL) and (AMNL) are contrasted with using only controls with normal likelihood ("CONT").

5. CONCLUSIONS

5.1 Nonparametric Regression Method for the Secondary Analysis in Case-Control Studies

Primarily, in the second Chapter, we have considered the case of nonparametrically estimating $E(Y|X)$ when no assumptions about the distribution of Y given X are made, including homoscedasticity. First we describe methodology in the rare case that the disease rate in the population, π_1, is known. We describe methodology in the far more common case that π_1 is unknown. In this common case, our simulations show conclusively that our tilted kernel estimator is the more efficient.

We considered the case that the disease rate in the population is unknown, and when one is willing to specify a distribution for Y given X up to a function $\mu(X)$ and other parameters, using a local likelihood method along with profiling methods. We displayed the method for when Y is binary with mean $H\{\mu(X)\}$. However, we emphasized two important points: (a) such methods are not consistent if the parametric model is misspecified; and (b) it is likely that the logistic intercept θ_0 will be very difficult to estimate numerically, and a rare disease approximation will improve computational performance (since it eliminates θ_0) while entailing little if any bias.

Ours is the first work to consider nonparametric regression in the secondary analysis of case-control studies. We have focused on the case of scalar X, and discovered a tilted kernel approach for estimation. With this tilted kernel function, extensions to multivariate X are surely possible, including multivariate kernel regression (Ruppert and Wand, 1994), additive models, etc.

5.2 Semiparametric Regression Method for the Heteroscedastic Single-Index Model

In the third Chapter we attempt to model both the mean and variance function by single-index model, a popular multivariate semiparametric model with a powerful dimension reduction quality. To our best knowledge, one existing work by Lian, et al. (2014) addresses the issue by local linear kernel approach. In the numerical examples, we found that the method is quite sensitive to the starting values and thus can have serious issues in practical applications. We have an entirely different approach based on random effects B-spline methodology of mean and variance function, with flexible truncated Dirichlet process mixture of normals for the regression errors. We use Bayesian computation to fit, because the structure makes such computation simple and straightforward. In simulation, we found that the precision of the estimates largely increases under our method compared to that of the kernel approach. Even with initial values selected by simple methods like ordinary ridge regression and principal hessian direction method (Li, 1991), our method yields consistent and efficient estimations in 5000 MCMC steps which takes about 15 mins in R using the Normal method and about 30 mins using the mixture of Normals. However, whether single-index model can be applicable to large scale problems where $n < p$ needs future study. An important extension of the current work and the subject of an ongoing research project is to find an integrated model which can effectively do variable selection of pertinent covariates to do single-index regression.

REFERENCES

Antoniadis, A., Grégoire, G. and McKeague, I. W. (2004). Bayesian estimation in single-index models. *Statistica Sinica*, 14, 1147-1164.

Bellman, R. (1961). Adaptive control process. *Princeton University Press*, 94.

Berkhof, J., Mechelen, I. and Gelman, A. (2003). A Bayesian approach to the selection and testing of mixture models. *Statistica Sinica*, 13, 423 - 442.

Bickel, P.J. (1978). Using residuals robustly I: tests for heteroscedasticity, nonlinearity. *Annals of Statistics*, 6, 266-291.

Boor, D. C. (2001). A practical guide to splines. *Springer*

Box, G. E. P. and Hill, W. J. (1974), Correcting inhomogeneity of variance with power transformation weighting, *Technometrics*, 16, 385 - 389.

Box, G. E. P. and Meyer, R. D. (1986). Dispersion effects from fractional designs. *Technometrics*, 28, 19-28.

Box, G. E. P. and Ramirez, J. (1986). Studies in quality improvement: signal to noise ratios, performance criteria and statistical analysis: Part II. Report no. 12.

Bush, C. E. and MacEachern, S. N. (1996). A semiparamteric Bayesian model for randomised block designs. *Biometrika*, 83, 275-285.

Cai, T. and Wang, L. (2008). Adaptive variance function estimation in heteroscedastic nonparametric regression, *Annals of Statistics*, 36, 2025 - 2054.

Carroll, R. J. (1982). Adapting for heteroscedasticity in linear models. *Annals of Statistics*, 10, 1224-1233.

Carroll, R. J. and Härdle, W. (1989). Second order effects in semiparametric weighted least squares regression. *Statistics*, 20, 179-186.

Carroll, R. J. and Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *Annals of Statistics*, 10, 429-441.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92, 447-489.

Carroll, R. J., Wang, C. Y. and Wang, S. (1995). Prospective analysis of logistic case-control studies. *Journal of the American Statistical Association*, 90, 157-169.

Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation in case-control studies of gene-environment interactions. *Biometrika*, 92, 399-418.

Chatterjee, N., Kalaylioglu, Z. and Carroll, R. J. (2005). A new paradigm of conditional-likelihoods for exploiting gene-environment independence in family based case-control studies. *Genetic Epidemiology,* 28, 138 - 156.

Chatterjee, N., Chen, Y.-H., Luo, S. and Carroll, R. J. (2009). Analysis of case-control association studies: SNPs, Imputation and Haplotypes. *Statistical Science*, 24, 489-502.

Chen, Y.-H., Carroll, R. J. and Chatterjee, N. (2008). Retrospective analysis of haplotype-based case-control studies under a flexible model for gene-environment association. *Biostatistics*, 9, 81-99.

Chen, Y.-H., Chatterjee, N. and Carroll, R. J. (2009). Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *Journal of*

the American Statistical Association, 104, 220-233.

Chen, Y.-H., Chatterjee, N. and Carroll, R. J. (2013). Using shared genetic controls in studies of gene-environment interactions. *Biometrika*, 100, 319-338.

Cook, R. D., and Li, B. (2002). Dimension reduction for conditional mean in regression. *Annals of Statistics*, 30, 455-474.

Davidian, M. and Carroll, R. J. (1987). Variance function estimation. *Journal of the American Statistical Association*, 82, 1079-1091.

Epstein, M. and Satten, G. A. (2003). Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics*, 73, 1316-1329.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, 11, 89 -121.

Escober, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90, 577-588.

Ferguson, T. F. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1, 209-230.

Fuller, W. A. and Rao, J. N. K. (1978). Estimation for a linear regression model with unknown diagonal covariance matrix. *Annals of Statistics*, 6, 1149-1158.

Gelfand, A. E., Kottas, A. and MacEachern, S.N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100, 1021 - 1035.

Griffin, J. E. and Steel, M. F. J. (2010). Bayesian nonparametric modelling with the dirichlet process regression smoother. *Statistica Sinica*, 20, 1507-1527.

Hall, P. and Carroll, R. J. (1989). Variance function estimation in regression: The effect of estimating the mean. *Journal of the Royal Statistical Society, Series B*, 51, 3 14

Härdle, W. and Stoker, M. T. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84, 986 - 995.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single index models. *Annals of Statistics*, 21, 157-178.

Hu, Y. J., Lin, D. Y. and Zeng, D. (2010). A general framework for studying genetic effects and geneenvironment interactions with missing data. *Biostatistics*, 11, 583-598.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 71-120.

Ishwaran, H. and James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96, 161-173.

Jiang, Y., Scott, A. J. and Wild, C. J. (2006). Secondary analysis of case-control data. *Statistics in Medicine*, 25, 1323-1339.

Kwee, L. C., Epstein, M. P., Manatunga, A. K., Duncan, R., Allen, A. S. and Satten, G. A. (2007). Simple methods for assessing haplotype-environment interactions in case-only and case-control studies. *Genetic Epidemiology*, 31, 75-90.

Keilegom, I. V. and Wang L. (2010). Semiparametric modeling and estimation of heteroscedasticity in regression analysis of cross-sectional data. *Electronic Journal of Statistics*, 4, 133-160.

Leslie, D. S., Kohn, R. and Nott, D.J. (2007). A general approach to heteroscedastic linear regression. *Statistical Computation*, 17, 131-146.

Li, K. C. (1991). Sliced Inverse Regression for dimension reduction (with discussion), *Journal of the American Statistical Association*, 86, 316-342.

Li, K. C. (1992). On principal hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *Journal of the American Statistical Association*, 87, 1025-1039.

Li, P. and Chen, J. (2010). Testing the order of a finite mixture. *Journal of the American Statistical Association*,105, 1084 -1092.

Li, H., Gail, M. H., Berndt, S. and Chatterjee, N. (2010). Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies. *Genetic Epidemiology*, 34, 427-433.

Lian, H., Liang, H. and Carroll, R. J. (2014). Variance function partially linear single-index models. *Journal of the Royal Statistical Society, Series B*, to appear.

Lin, D. Y.and Zeng D. (2006). Likelihood-based inference on haplotype effects in genetic association studies (with discussion). *Journal of the American Statistical Association*, 101, 89-118.

Lin, D. Y. and Zeng, D. (2009). Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology*, 33, 256-265.

Liu, M. C. and Taylor, R. L. (1989). A consistent nonparametric density estimator for the deconvolution problem. *Canadian Journal of Statistics*, 17, 427-438.

Lo, A.Y. (1984). On a class of bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics*, 12, 351 - 357.

Ma, Y., Chiou, J., M and Wang, N. (2006). Efficient semiparametric estimator for heteroscedastic partially linear models. *Biometrika*, 93, 75-84.

Ma, Y. and Zhu, L. (2012). A semiparametric approache to dimension reduction. *Journal of the American Statistical Association*, 107, 168-179.

Ma, Y. and Zhu, L. (2013). Efficient Estimation in Sufficient Dimension Reduction. *Annals of Statistics*, 41, 250-268.

McKeague, I. W., and Tighiouart, M. (2000). Bayesian Estimators for Conditional Hazard Functions. *Biometrics*, 56, 1007-1015.

Modan, M. D., Hartge, P., Hirsh-Yechezkel, G., Chetrit, A., Lubin, F., Beller, U., Ben-Baruch, G., Fishman, A., Menczer, J., Struewing, J. P., Tucker, M. A. and Wacholder, S. for the National Israel Ovarian Cancer Study Group (2001). Parity, oral contraceptives and the risk of ovarian cancer among carriers and noncarriers of a BRCA1 or BRCA2 mutation. *The New England Journal of Medicine*, 345, 235-240.

Monsees, G., Tamimi, R. and Kraft, P. (2009). Genomewide association scans for secondary traits using case-control samples. *Genetic Epidemiology*, 33, 717728.

Nagelkerke, N.J.D., Moses, S., Plummer, F.A., Brunham, R.C., Fish, D. (1995). Logistic regression in case-control studies: the effect of using independent as dependent variables. *Statistics in Medicine*, 14, 769 775.

Pelenis, J. (2014). Semiparametric Bayesian regression. *Journal of Econometrics*, 178, 624-638.

Powell, L. J., Stock, H. J. and Stocker, M. T. (1989). Semiparametric estimation of index coefficient. *Econometrica*, 57, 1403-1430.

Piegorsch, W. W., Weinberg, C. R. and Taylor, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population based case-control studies. *Statistics in Medicine*, 13, 153-162.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403-411.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11, 735-757.

Ruppert, D and Wand, M. P. (1997). Multivariate locally weighted least square regression. *Annals of Statistics*, 22, 1346 - 1370.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An Effective Bandwidth Selector for Local Least Squares Regression. *Journal of the American Statistical Association*, 90, 1257-1270.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, 639-650.

Teschendorff, A. E. and Widschwendter, M. (2012). Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics*, 28, 1487-1494.

Wang, B. H. (2009). Bayesian estimation and variable selection for single index models. *Computational Statistics & Data Analysis*, 53, 2617-2627.

Wang, C. Y., Wang, S. and Carroll, R. J. (1997). Estimation in choice-based sampling with measurement error and bootstrap analysis. *Journal of Econometrics*,

77, 65-86.

Wei, J., Carroll, R.J., Mü, U., Van Keilegom, I. and Chatterjee, N. (2013). Locally efficient estimation for homoscedastic regression in the secondary analysis of case-control data. *Journal of Royal Statistical Society, Series B*, 75, 185-206.

West, M., Müller, P. and Escober, M. D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. *John Wiley*, 363-386.

Xia, Y. (2002). Single-index volatility models and estimation. *Statistica Sinica*, 12, 785-799.

Xia, Y. and Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis*, 97, 1162-1184.

Yang, Q., Khoury, M. J. and Flanders, W. D. (1997). Sample size requirements in case-only designs to detect gene-environment interaction. *American Journal of Epidemiology*, 146, 713720.

Yau, P. and Kohn, R. (2003). Estimation and variable selection in nonparametric heteroscedastic regression. *Statistics and Computing*, 13, 191-208.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimators for partially linear single index models. *Journal of the American Statistical Association*, 97, 1042 - 1054.

Zhao, L. P., Li, S. S. and Khalid N. (2003). A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *American Journal of Human Genetics*, 72, 1231-1250.

Zhu, L. X., and Fang, K. T. (1996). Asymptotics for kernel estimation of Sliced Inverse Regression, *Annals of Statistics*, 3, 1053-1068.

APPENDIX A

FIRST APPENDIX

A.1 Notation and Supporting Lemmas

A.1.1 Notation

In this Section we introduce notation that is needed for deriving the main results. Superscript $^{(1)}$ refers to a first derivative, $^{(2)}$ refers to a second derivative, etc. We assume that $f_{XY}(\cdot)$, $f_X(\cdot)$ and $f_{X,cont}(\cdot)$ is twice continuously and boundedly differentiable with respect to x.

Define

$$\begin{aligned} \mu(x_0) &= E\{Y|X=x_0\},\ \ \mu_{cont}(x_0) = E\{Y|D=0\}, \\ \mathcal{K}^{(j)}(y,\Omega,x) &= \partial^{(j)}\mathcal{K}(y,x,\Omega)/\partial x^j; \end{aligned}$$

Write

$$\begin{aligned} H(d=0|y,x,\theta_0,\theta_1) &= \mathcal{K}_{pop}(y,x_0,\Omega,\theta_0)/\mathcal{K}(y,x_0,\Omega) \\ &= \{1+\exp(\theta_0+m(y,x_0,\theta_1)\}^{-1}; \\ \mu_{\text{cont}}(x_0) &= \frac{\int y f_{YX}(y,x_0)H(d=0|y,x,\theta_0,\theta_1)dy}{\int f_{YX}(y,x_0)H(d=0|y,x,\theta_0,\theta_1)dy}, \end{aligned}$$

Define

$$\begin{aligned}
\widehat{m}_h(x_0) &= C_{1n}(x_0)/C_{0n}(x_0), \\
\text{for p = 0,1} \quad C_{pn}(x_0) &= n^{-1}\sum_{i=1}^{n} Y_i^p G_0(X_i, \Omega, x_0), \\
G_0(X_i, \Omega, x_0) &= K_h(X_i - x_0)/\Lambda_n(Y_i, x_0, h, \Omega), \\
\Lambda_n(Y_i, x_0, h, \Omega) &= n_0^{-1}\sum_{j=1}^{n}(1 - D_j)K_h(X_j - x_0)\mathcal{K}(y, X_j, \Omega),
\end{aligned}$$

$$\begin{aligned}
\text{for p=0,1} \quad M_{np}(y, x, x_0, \Omega) &= n^{-1}\sum_{i=1}^{n} Y_i^p K_h(X_i - x_0)/a_2(Y_i, \Omega, x_0), \\
D_{np}(y, x, x_0, \Omega) &= n^{-1}\sum_{i=1}^{n} Y_i^p K_h(X_i - x_0)a_3(y, \Omega, x)/a_2^2(Y_i, \Omega, x_0), \\
a_2(y, x_0, \Omega) &= f_{X,cont}(x_0)\mathcal{K}(Y_i, \Omega, x_0), \\
a_3(y, \Omega, x_0) &= \mathcal{K}^{(1)}(y, \Omega, x_0)f^{(1)}_{X,cont}(x_0) + (1/2)K^{(2)}(y, \Omega, x_0)f_{X,cont} \\
&\quad + 1/2\mathcal{K}(y, x_0, \Omega)f^{(2)}_{X,cont}(x_0),
\end{aligned}$$

For p=0,1, also define

$$\begin{aligned}
B_p(x_0, \Omega, \theta_0) &= \int \frac{a_3(y, \Omega, x_0)}{a_2^2(y, \Omega, x_0)} y^p \{f^{(1)}_{XY}(y, x_0)\mathcal{K}^{(1)}_{pop}(y, x_0, \Omega, \theta_0) \\
&\quad + 1/2 f_{XY}(y, x_0)\mathcal{K}^{(2)}_{pop}(y, x_0, \Omega, \theta_0) + 1/2\mathcal{K}(y, x_0, \Omega)f^{(2)}_{XY}(y, x_0) + o(h^2)\}dy, \\
R_p(x_0, \Omega, \theta_0) &= \int y^p a_3(y, \Omega, x_0)/a_2^2(y, \Omega, x_0) f_{XY}(y, x_0)\mathcal{K}_{pop}(y, x_0, \Omega, \theta_0)dy, \\
M_p(x_0) &= \int y^p \frac{f_{YX}(y, x_0)}{f_{X,cont(x_0)}} H(d = 0|y, x, \theta_0, \theta_1)dy,
\end{aligned}$$

Let

$$
\begin{aligned}
U_0(x_0,\Omega,\theta_0) &= c_2 \int K_{pop}(y,x_0,\Omega,\theta_0)/a^2(y,x_0,\Omega) f_{YX}(y,x_0)dy, \\
U_1(x_0,\Omega,\theta_0) &= c_2 \int y^2 K_{pop}(y,x_0,\Omega,\theta_0)/a^2(y,x_0,\Omega) f_{YX}(y,x_0)dy, \\
U_2(x_0,\Omega,\theta_0) &= c_2 \int y K_{pop}(y,x_0,\Omega,\theta_0)/a^2(y,x_0,\Omega) f_{YX}(y,x_0)dy, \\
U(x_0,\Omega,\theta_0) &= \frac{U_0(x_0,\Omega,\theta_0)M_1^2(x_0)}{M_0(x_0)} + \frac{U_1(x_0,\Omega,\theta_0)}{M_0^2(x_0)} - \frac{2U_2(x_0,\Omega,\theta_0)M_1(x_0)}{M_0^3(x_0)} \\
W(x_0,\Omega,\theta_0) &= M_0(x_0)^{-1}\{c_1 R_1(x_0,\Omega,\theta_0) - R_0(x_0,\Omega,\theta_0)M_1(x_0)/M_0^2(x_0)\} \\
b_0 &= \left\{\pi_0 U(x_0,\Omega,\theta_0)/4W^2(x_0,\Omega,\theta_0)\right\}^{1/5}.
\end{aligned}
$$

A.1.2 Lemma 1

We first state a lemma that will be used repeatedly in the development of the asymptotic theory. Recall that the retrospective likelihood for Y, X given $D = d$ is $\mathrm{pr}(Y, X|D = d) = \mathrm{pr}(D = d|Y, X)\mathrm{pr}(Y, X)/\mathrm{pr}(D = d)$.

Lemma 1 *Let "E_{cc}" denote expectation under the case-control sampling design, i.e., conditional on D. Under the case-control sampling design, conditioned on the disease status, for any measurable function $Q(Y, X)$ of data (D, Y, X), the retrospective expectation is given by*

$$
\begin{aligned}
E_{cc}\{n^{-1}\sum_{i=1}^{n} Q(Y_i,X_i)\} &= n^{-1}\sum_{i=1}^{n} E\{Q(Y_i,X_i)|D_i = d_i\} \\
&= \{n_0/(n\pi_0)\}\int\int f_{XY}(x,y)Q(y,x)\mathcal{K}_{pop}(y,x,\Omega)dydx
\end{aligned}
$$

Proof. Using the notation in Section A.1.1 we deduce

$$\begin{aligned}
E_{cc}\{n^{-1}\sum_{i=1}^{n} Q(Y_i, X_i)\} &= E\{n^{-1}\sum_{i=1}^{n} Q(Y_i, X_i)|D = D_i\} \\
&= \sum_{d=0}^{1}(n_d/n)E\{Q(y,x)|D=d\} \\
&= \sum_{d=0}^{1}\{n_d/(n\pi_d)\}\int\int Q(y,x)f(d|y,x)f_{Y,X}(x,y)dxdy \\
&= (1/n)\int\int Q(y,x)f_{YX}(y,x)\frac{n_0/\pi_0 + n_1/\pi_1 \exp\{\theta_0 + m(y,x,\theta_1)\}}{1+\exp\{\theta_0 + m(y,x,\theta_1)\}}dxdy \\
&= \{n_0/(n\pi_0)\}\int\int f_{YX}(x,y)Q(y,x)\frac{1+\exp\{\kappa + m(y,x,\theta_1)\}}{1+\exp\{\theta_0 + m(y,x,\theta_1)\}}dydx \\
&= \{n_0/(n\pi_0)\}\int\int f_{YX}(x,y)Q(y,x)\mathcal{K}_{pop}(y,x,\Omega,\theta_0)dydx,
\end{aligned}$$

as claimed.

A.1.3 Supplementary Lemmas

Here we provide supporting lemmas which supplement the derivation of the asymptotic theory of $\widehat{m}_h(x_0)$ defined at (2.10). Recall the notation in appendix A.1.1 where we defined $\widehat{m}_h(x_0) = C_{1n}(x_0)/C_{0n}(x_0)$. In this section, our aim is to show $C_{1n}(x_0)/C_{0n}(x_0) \to \mu_{cont}(x_0)$ in probability.

Lemma 2 *Recall the definition of* $\Lambda_n(y,x,h,\Omega)$ *at (2.9).*

$$\begin{aligned}
\Lambda_n(Y_i, x_0, h, \Omega) &= n_0^{-1}\sum_{j=1}^{n}(1-D_j)K_h(X_j - x_0)\mathcal{K}(y, X_j, \Omega) \\
&= f_{X,cont}(x_0)\mathcal{K}(y,x,\Omega) + h^2 c_1 a_3(y,\Omega,x_0) + \mathcal{O}_p\{(n_0h)^{-1/2}\} + \\
&\quad o_p\{h^2 + (n_0h)^{-1/2}\}.
\end{aligned}$$

Proof. Define $z = (x - x_0)/h$, using the Taylor series expansion w.r.t x_0 and recall $\int K(z) = 1$, $\int zK(z) = 0$ and $\int z^2K(z) = c_1$, we obtain the expectation and variance

among the controls

$$
\begin{aligned}
E\{\Lambda_n(Y_i, x_0, h, \Omega)|D=0\} &= \int K_h(x-x_0)\mathcal{K}(y,x,\Omega)f_{X,cont}(x)dx \\
&= (1/h)\int K(z)\mathcal{K}(y, x_0+zh, \Omega)f_{X,cont}(x_0+zh)dx \\
&= \int K(z)\big\{\mathcal{K}(y,x_0,\Omega) + zh\mathcal{K}^{(1)}(y,x_0,\Omega) + z^2h^2\mathcal{K}^{(2)}(y,x_0,\Omega) + o_p(h^2)\big\} \\
&\qquad \times\big\{f_{X,cont}(x_0) + zhf^{(1)}_{X,cont}(x_0) + z^2h^2f^{(2)}_{X,cont}(x_0) + o_p(h^2)\big\}dx \\
&= f_{X,cont}(x_0)\mathcal{K}(y,x_0,\Omega) + h^2c_1\{\mathcal{K}^{(1)}(y,\Omega,x_0)f^{(1)}_{X,cont}(x_0) \\
&\qquad +(1/2)\mathcal{K}^{(2)}(y,\Omega,x_0)f_{X,cont} + (1/2)\mathcal{K}(y,x_0,\Omega)f^{(2)}_{X,cont}(x_0)\} + o_p(h^2);
\end{aligned}
$$

$$
\begin{aligned}
&\mathrm{var}\{n_0^{-1}\sum_{i=1}^{n}(1-D_j)K_h(X_j-x_0)\mathcal{K}(y,X_j,\Omega)\} \\
&\qquad \le n_0^{-1}\int K_h^2(X-x_0)\mathcal{K}^2(y,x,\Omega)f_{X,cont}(x)dx \\
&\qquad = (n_0h)^{-1}\int K^2(z)\mathcal{K}^2(y,x_0+zh,\Omega)f_{x,cont}(x_0+zh)dz \\
&\qquad = \mathcal{O}_p(n_0h)^{-1} + o_p(n_0h)^{-1}.
\end{aligned}
$$

Hence

$$
\begin{aligned}
\Lambda_n(Y_i,x_0,h,\Omega) &= f_{X,cont}(x_0)\mathcal{K}(y,x,\Omega) + h^2c_1a_3(y,\Omega,x_0) \\
&\qquad +\mathcal{O}_p\{(n_0h)^{-1/2}\} + o_p\{h^2+(n_0h)^{-1/2}\} \\
&= a_2(y,x_0,\Omega) + h^2c_1a_3(y,\Omega,x_0) + \mathcal{O}_p\{(n_0h)^{-1/2}\} + o_p\{h^2+(n_0h)^{-1/2}.
\end{aligned}
$$

Lemma 3

$$
\begin{aligned}
D_{np}(y,x_0,\Omega) &= n_0/(n\pi_0)\int y^p a_3(y,x_0,\Omega)/a_2^2(y,x_0,\Omega)f_{XY}(y,x_0)\mathcal{K}_{pop}(y,x_0,\Omega,\theta_0)dy \\
&\quad + \mathcal{O}_p(h^2) + \mathcal{O}_p(nh)^{-1/2} + o_p\{(nh)^{-1/2}+h^2\},
\end{aligned}
$$

and

$$\begin{aligned} M_{np}(y,x_0,\Omega) &= \frac{n_0}{n\pi_0 f_{X,cont}(x_0)}\int y^p f_{YX}(y,x_0)H(d=0|y,x,\theta_0,\theta_1)dy + \mathcal{O}_p(h^2) \\ &+ \mathcal{O}_p(nh)^{-1/2} + o_p\{(nh)^{-1/2}+h^2\}. \end{aligned}$$

Proof. Using Lemma 1 and Taylor series expansion we obtain

$$\begin{aligned} E\big(D_{np}|D=d\big) &= E\{n^{-1}\sum_{i=1}^{n} Y_i^p \frac{K_h(X_i-x_0)a_3(y,\Omega,x_0)}{a_2^2(y,\Omega,x_0)}|D=D_i\} \\ &= n_0/(n\pi_0)\int\int y^p K_h(x-x_0)\frac{a_3(y,x_0,\Omega)}{a_2^2(y,x_0,\Omega)}f_{XY}(y,x)\mathcal{K}_{pop}(y,\Omega,x)dydx \\ &= n_0/(n\pi_0)\int y^p \frac{a_3(y,\Omega,x_0)}{a_2^2(y,\Omega,x_0)}f_{XY}(y,x_0)\mathcal{K}_{pop}(y,x_0,\Omega,\theta_0)dy \\ &+ h^2c_1\int y^p\frac{a_3(y,\Omega,x_0}{a_2^2(y,\Omega,x_0)}\{f_{XY}^{(1)}(y,x_0)\mathcal{K}_{pop}^{(1)}(y,x_0,\Omega,\theta_0)+ \\ &+ (1/2)f_{X,Y}(y,x_0)\mathcal{K}_{pop}^{(2)}(y,\Omega,x_0) \\ &+(1/2)\mathcal{K}_{pop}(y,x_0,\Omega,\theta_0)f_{XY}^{(2)}(y,x_0)+o(h^2)\}dy \\ &= n_0/(n\pi_0)R_p(x_0,\Omega,\theta_0)+h^2n_0/(n\pi_0)B_p(x_0,\Omega,\theta_0)+o_p(h^2). \end{aligned}$$

The case-control variance of D_{np} is

$$\begin{aligned} \mathrm{var}\big(D_{np}|D=d\big) &= n^{-2}\sum_{i=1}^{n}\mathrm{var}\{Y_i^p\frac{K_h(X_i-x_0)a_3(y,\Omega,x_0)}{a_2^2(y,\Omega,x_0)}|D=D_i\} \\ &\leq \mathcal{O}_p(nh)^{-1}+o_p\{(nh)^{-1}+h^2\}. \end{aligned}$$

Similarly we can prove that

$$\begin{aligned} E_{cc}\{M_{np}(y,x,\Omega)\} &= \frac{n_0}{n\pi_0 f_{X,cont}(x_0)}\int y^p f_{YX}(y,x_0)\frac{\mathcal{K}_{pop}(y,x_0,\Omega,\theta_0)}{\mathcal{K}(y,x_0,\Omega)}dy+\mathcal{O}_p(h^2)+o_p(h^2) \\ &= n_0/(n\pi_0)M_p(x_0)+\mathcal{O}_p(h^2)+o_p(h^2), \\ \mathrm{var}\{M_{np}(y,x,\Omega)\} &\leq \mathcal{O}_p(nh)^{-1}+o_p\{(nh)^{-1}+h^2\}. \end{aligned}$$

Lemma 4

$$\begin{aligned} C_{pn} &= M_p(x_0) + h^2 c_1 \frac{n_0}{n\pi_0} B_p(x_0, \Omega, \theta_0) + h^2 \frac{n_0}{n\pi_0} R_p(x_0, \Omega, \theta_0) \\ &\quad + \mathcal{O}_p\{(n_0 h)^{-1/2}\} + o_p\{(nh)^{-1/2} + h^2\} \end{aligned}$$

Proof. From the notation in Appendix A.1.1, we know that

$$C_{pn}(x_0) = n^{-1} \sum_{i=1}^{n} \frac{Y_i^p K_h(X_i - x_0)}{\Lambda_n(Y_i, x_0, h, \Omega)},$$

Using Lemma 2, and defining $\mathcal{V}_n = \mathcal{O}_p\{(n_0 h)^{-1/2}\} + o_p\{h^2 + (n_0 h)^{-1/2}\}$,we expanded the denominator $\Lambda_n(Y_i, x_0, h, \Omega)$ to obtain

$$C_{pn}(x_0) = n^{-1} \sum_{i=1}^{n} \frac{Y_i^p K_h(X_i - x_0)}{f_{X,cont}(x_0)\mathcal{K}(Y_i, x, \Omega) + h^2 c_1 a_3(Y_i, \Omega, x_0) + \mathcal{V}_n},$$

By the definition of $a_2(y, x_0, \Omega)$, we obtain

$$\begin{aligned} C_{pn}(x_0) &= n^{-1} \sum_{i=1}^{n} Y_i^p \frac{K_h(X_i - x_0)}{a_2(Y_i, \Omega, x_0)} \\ &\quad - c_1 h^2 n^{-1} \sum_{i=1}^{n} Y_i^p \frac{K_h(X_i - x_0) a_3(y, \Omega, x)}{a_2^2(Y_i, \Omega, x_0)} + \mathcal{V}_n \\ &= M_{np}(y, x, x_0, \Omega) - c_1 h^2 D_{np}(y, x, x_0, \Omega) + \mathcal{V}_n. \end{aligned}$$

Now we use Lemma 3 for $M_{np}(y, x, x_0, \Omega)$ to obtain

$$C_{pn} = \frac{n_0}{n\pi_0} M_p(x_0) + h^2 c_1 \frac{n_0}{n\pi_0} R_p(x_0, \Omega, \theta_0) + h^4 \frac{n_0}{n\pi_0} B_p(x_0, \Omega, \theta_0) + \mathcal{V}_n.$$

A.2 Asymptotic Theory

B.1 Proof of Theorem 1

We start with finding the case control expectation of the denominator of (2.6) as

$$
\begin{aligned}
&E\big\{n^{-1}\sum_{i=1}^{n} K_h(X_i - x_0)/\Lambda_{pop}(Y_i, x_0, h, \Omega, \theta_0)|D_i = d_i\big\} \\
&\quad = \{n_0/(n\pi_0)\}\int\int \frac{Y_i K_h(X_i - x_0)}{\Lambda_{pop}(Y_i, x_0, h, \Omega, \theta_0)} f_{XY}(x, y)\mathcal{K}_{pop}(y, x, \Omega, \theta_0)dydx \\
&\quad = \{n_0/(n\pi_0)\}\int \frac{\int K_h(x - x_0)\mathcal{K}_{pop}(y, x, \Omega, \theta_0)f_{XY}(x, y)dx}{\int K_h(v - x_0)\mathcal{K}_{pop}(y, v, \Omega, \theta_0)f_X(v)dv}dy,
\end{aligned}
$$

By Taylor series expansion of $f_{XY}(x, y)$ and $\mathcal{K}_{pop}(y, x, \Omega, \theta_0)$ with respect to x about x_0 we have

$$
\begin{aligned}
&\int f_{XY}(x, y)K_h(x - x_0)\mathcal{K}_{pop}(y, x, \Omega, \theta_0)dx \\
&\quad = f_{XY}(x_0, y)\mathcal{K}_{pop}(y, x_0, \Omega, \theta_0) + c_1 h^2\{f^{(1)}_{XY}(x_0, y)\mathcal{K}^{(1)}_{pop}(y, x_0, \Omega, \theta_0) \\
&\qquad +(1/2)f^{(2)}_{XY}(x_0, y)\mathcal{K}_{pop}(y, x_0, \Omega, \theta_0) + f_{XY}(x_0, y)\mathcal{K}^{(2)}_{pop}(y, x_0, \Omega, \theta_0)\} + o(h^2),
\end{aligned}
$$

and

$$
\begin{aligned}
&\int K_h(v - x_0)\mathcal{K}_{pop}(y, v, \Omega)f_X(v)dv \\
&\quad = f_X(x_0)\mathcal{K}_{pop}(y, x_0, \Omega, \theta_0) + h^2 c_1 f^{(1)}_X(x_0)\mathcal{K}^{(1)}_{pop}(y, x_0, \Omega, \theta_0) \\
&\qquad +(1/2)h^2 c_1\{f^{(2)}_X(x_0)\mathcal{K}_{pop}(y, x_0, \Omega, \theta_0) + f_X(x_0)\mathcal{K}^{(2)}_{pop}(y, x_0, \Omega, \theta_0)\} + o(h^2).
\end{aligned}
$$

Further, with some algebra we can show that as $h \to 0, n \to \infty$ and $n_0/n \to c$

$$\begin{aligned} & E\big\{n^{-1}\sum_{i=1}^{n} K_h(X_i - x_0)/\Lambda_{pop}(Y_i, x_0, h, \Omega, \theta_0)|D_i = d_i\big\} \\ &= \frac{n_0}{n\pi_0 f_X(x_0)} \int f_{XY}(x_0, y)\frac{\mathcal{K}_{pop}(y, x_0, \Omega, \theta_0)}{\mathcal{K}_{pop}(y, x_0, \Omega, \theta_0)}dy + \ \mathcal{O}(h^2) + o(h^2) \\ &= \frac{n_0}{n\pi_0 f_X(x_0)} \int f_{XY}(x_0, y)dy + \mathcal{O}(h^2) + o(h^2). \end{aligned}$$

Similarly we have

$$\begin{aligned} \lim_{h\to 0} E\big\{n^{-1}\sum_{i=1}^{n} Y_i K_h(X_i - x_0)/\Lambda_{pop}(Y_i, x_0, h, \Omega, \theta_0)|D_i = d_i\big\} \\ = \frac{n_0}{n\pi_0 f_X(x_0)} \int y f_{XY}(x_0, y)dy \end{aligned}$$

Thus, when $h \to 0$ and $nh \to \infty$

$$\widehat{M}_h(x_0) = \frac{\int y f_{XY}(x_0, y)dy}{\int f_{XY}(x_0, y)dy} + o_p(1) = \ \mu(x_0) + o_p(1).$$

B.2 Proof of Theorem 2

Recall the notation in appendix A.1.1 where we defined $\widehat{m}_h(x_0) = C_{1n}(x_0)/C_{0n}(x_0)$ and the definition $\mathcal{V}_n = \mathcal{O}_p\{(n_0h)^{-1/2}\} + o_p\{h^2 + (n_0h)^{-1/2}\}$. We first use the Taylor series expansion of C_{1n}/C_{0n} to derive the consistency of $\widehat{m}_h(x_0)$ to $\mu_{\text{cont}}(x_0)$. Define

$\mathcal{A} = n_0/(n\pi_0)$.

$$
\begin{aligned}
&\widehat{m}_h(x_0) \\
&= \frac{C_{1n}}{C_{0n}} \\
&= \frac{\mathcal{A}M_1(x_0) + h^2 c_1 \mathcal{A} R_1(x_0,\Omega,\theta_0) + h^4 \mathcal{A} B_1(x_0,\Omega,\theta_0) + \mathcal{V}_n}{\mathcal{A}M_0(x_0) + h^2 c_1 \mathcal{A} R_0(x_0,\Omega,\theta_0) + h^4 \mathcal{A} B_0(x_0,\Omega,\theta_0) + \mathcal{V}_n} \\
&= M_1(x_0) + h^2 R_1(x_0,\Omega,\theta_0) + \mathcal{V}_n \\
&\qquad \times \; [\frac{1}{M_0(x_0)} - \frac{h^2}{M_0^2(x_0)} R_0(x_0,\Omega,\theta_0) + \mathcal{V}_n] \\
&= \frac{M_1(x_0)}{M_0(x_0)} + h^2 M_0(x_0)^{-1}\{c_1 R_1(x_0,\Omega,\theta_0) - \frac{R_0(x_0,\Omega,\theta_0) M_1(x_0)}{M_0^2(x_0)}\} + \mathcal{V}_n \\
&= \mu_{\text{cont}}(x_0) + \mathcal{O}_p(h^2) + \mathcal{V}_n.
\end{aligned}
$$

We thus obtain the asymptotic bias of $\widehat{m}_h(x_0)$ as

$$
\begin{aligned}
E\{\widehat{m}_h(x_0) - \mu_{\text{cont}}(x_0)\} &= h^2/M_0(x_0)\{c_1 R_1(x_0,\Omega,\theta_0) - \frac{R_0(x_0,\Omega,\theta_0) M_1(x_0)}{M_0^2(x_0)}\} + \mathcal{V}_n \\
&= h^2 W(x_0,\Omega,\theta_0) + \mathcal{V}_n.
\end{aligned}
$$

Next, we use Taylor series so that $\text{var}\big(C_{1n}/C_{0n}\big) \approx \alpha_1^2/\alpha_0^4 \text{var}(C_{0n}) + 1/\alpha_0^2 \text{var}(C_{1n}) - 2\alpha_1/\alpha_0^3 \text{cov}(C_{1n}, C_{0n})$, where $\alpha_1 = E(C_{1n})$ and $\alpha_0 = E(C_{0n})$. However, in our study both expectation and variance are take in case-control framework. So for p = 0,1 we will consider $E(C_{pn}|D=d) = n_0/(n\pi_0) M_p(x_0)$ and we deduce the variance

$$
\begin{aligned}
\text{var}\big(C_{1n}|D=d\big) &= \text{var}\left\{ \frac{n^{-1}\sum_{i=1}^n Y_i K_h(x_i - x_0)}{A_n(y,x,\Omega)} \middle| D = d_i \right\} \\
&= \text{var}\left\{ n^{-1} \sum_{i=1}^n \frac{Y_i K_h(x_i - x_0)}{a_2(Y_i, x_0, \Omega) + \mathcal{V}_n} \middle| D = d_i \right\} \\
&= n^{-1} \text{var}\left\{ \frac{Y_i K_h(x_i - x_0)}{a_2(Y_i, x_0, \Omega) + \mathcal{V}_n} \middle| D = d_i \right\},
\end{aligned}
$$

Consider

$$\begin{aligned}\mathrm{var}\left\{\frac{YK_h(x_i-x_0)}{a_2(Y,x_0,\Omega)}\Big|D=d\right\} &= E\left[\left\{\frac{YK_h(X-x_0)}{a_2(Y,x_0,\Omega)}|D=d\right\}^2\right] \\ &\quad - \left[E\left\{\frac{YK_h(X-x_0)}{a_2(Y,x_0,\Omega)}|D=d\right\}\right]^2,\end{aligned}$$

It is easy to see that

$$E\left[\left\{\frac{YK_h(X-x_0)}{a_2(Y,x_0,\Omega)}|D=d\right\}^2\right] = \frac{n_0}{nh\pi_0}U_1(x_0,\Omega,\theta_0)+o(\frac{h}{nh}),$$

where $U_1(x_0,\Omega,\theta_0) = c_2\int y^2K_{pop}(y,x_0,\Omega,\theta_0)/a^2(y,x_0,\Omega)f_{YX}(y,x_0)dy$ and $\int K^2(z)dz = c_2$. Hence,

$$\mathrm{var}\big(C_{1n}|D=d\big) = \frac{n_0}{n^2h\pi_0}U_1(x_0,\Omega,\theta_0) - n^{-1}\mathcal{A}^2M_1^2(x_0)+o(\frac{h}{nh}),$$

Arguing as before, we may deduce that

$$\begin{aligned}\mathrm{var}\big(C_{0n}|D=d\big) &= \frac{n_0}{n^2h\pi_0}U_0(x_0,\Omega,\theta_0) - n^{-1}\mathcal{A}^2M_0^2(x_0)+o(\frac{h}{nh}),\\ \mathrm{cov}\big(C_{1n},C_{0n}|D=d\big) &= \frac{n_0}{n^2h\pi_0}U_2(x_0,\Omega,\theta_0) - n^{-1}\mathcal{A}^2M_1(x_0)M_0(x_0)+o(\frac{h}{nh}),\end{aligned}$$

where $U_0(x_0,\Omega,\theta_0) = c_2\int K_{pop}(y,x_0,\Omega,\theta_0)/a^2(y,x_0,\Omega)f_{YX}(y,x_0)dy$ are defined in Appendix A.1.1. From the notation in Appendix A.1.1 for $U(x_0,\Omega,\theta_0)$, and simply by plugging in the Taylor series approximation of $\mathrm{var}\big(C_{1n}/C_{0n}\big)$ mentioned above , we obtain

$$\mathrm{var}\big\{\widehat{m}_h(x_0)\big\} = \frac{\pi_0}{n_0h}U(x_0,\Omega,\theta_0)+o(h/nh).$$

The approximate mean square error (MSE) of the estimator is thus

$$\begin{aligned} \mathrm{MSE}\{\widehat{m}_h(x_0)\} &= \mathrm{Bias}^2\{\widehat{m}_h(x_0)\} + \mathrm{var}\{\widehat{m}_h(x_0)\} \\ &\approx h^4 W^2(x_0, \Omega, \theta_0) + \frac{\pi_0}{n_0 h} U(x_0, \Omega, \theta_0). \end{aligned}$$

Minimizing the MSE obtained above we get the optimal bandwidth h_{opt} as proportional to $n_0^{-1/5}$, that is $h_{opt} = b_0 n_0^{-1/5}$, where $b_0 = \{\pi_0 U(x_0, \Omega, \theta_0)/4W^2(x_0, \Omega, \theta_0)\}^{1/5}$, as claimed.

B.3 Proof of Theorem 3

We next derive the asymptotic distribution of $\widehat{m}_h(x_0)$. First recall that $\widehat{m}_h(x_0) = C_{1n}/C_{0n}$. By Lemma 4, C_{0n} can be expressed as $\{n_0/(n\pi_0)\}M_p(x_0) + \mathcal{O}(h^2) + \mathcal{V}_n$. We define $Z_{ni} = Y_i K_h(x_i - x_0)/A_n(y_i, x_0, \Omega)$ and write

$$\begin{aligned} \widehat{m}_h(x_0) &= \pi_0 n_0 M_0(x_0)^{-1} \sum_{i=1}^{n} Y_i K_h(x_i - x_0)/A_n(Y_i, x_0, \Omega) + \mathcal{O}_p(h^2) + \mathcal{V}_n, \\ &= \pi_0 n_0 M_0(x_0)^{-1} \sum_{i=1}^{n} Z_{ni} + \mathcal{O}_p(h^2) + \mathcal{V}_n. \end{aligned}$$

so that $\{Z_{ni}\}$ is a triangular array of random variables. We will use Lyapounov's Central Limit Theorem for triangular arrays to derive the asymptotic distribution of our estimator. In particular, $s_n^2 = \mathrm{var}\{\widehat{m}_h(x_0)\} = \{\pi_0/(n_0 h)\}U(x_0, \Omega, \theta_0) + o(h/nh)$. The Lyapounov's condition holds if there exist $\delta > 0$ such that $s_n^{-(2+\delta)} \sum_{i=1}^{n} E\{|Z_{ni} -$

$E(Z_{ni})|\}^{2+\delta} \to 0$ for $n \to \infty$. Here, the condition is satisfied with $\delta = 1$ because

$$\begin{aligned}
\rho_{ni} = E_{cc}\Big\{|Z_{ni} - E(Z_{ni})|\Big\}^3 &\leq 8E_{cc}\big(Z_{ni}\big)^3 \\
&= \int \frac{8y^3K^3(u)}{h^2A_n(y,x_0)}\mathcal{K}_{pop}(y, x_0+uh)f_{YX}(y, x_0+uh)dydu \\
&\quad + \mathcal{O}(h^2) \\
&= \int y^3\frac{8c_3\mathcal{K}_{pop}(y, x_0+uh)f_{YX}(y, x_0+uh)}{h^2A_n(y,x_0)}dydu + \mathcal{O}(h^{-1}).
\end{aligned}$$

$$\text{Therefore,} \qquad \frac{\sum_{i=1}^n \rho_{ni}}{(s_n^2)^{3/2}} \leq \frac{\mathcal{O}(nh^{-2})}{\{\mathcal{O}(nh^{-1})\}^{3/2}} \to 0 \text{ if } nh \to \infty.$$

Hence, when $nh \to \infty$,

$$\begin{aligned}
(nh)^{1/2}\Big[\widehat{m}_h(x_0) - E\{\widehat{m}_h(x_0)\}\Big] &\to \text{Normal}\Big\{0, U(x_0, \Omega, \theta_0)\Big\} \\
(nh)^{1/2}\Big[\widehat{m}_h(x_0) - \mu_{\text{cont}}(x_0) - h^2W(x_0, \Omega, \theta_0)\Big] &\to \text{Normal}\Big\{0, U(x_0, \Omega, \theta_0)\Big\}.
\end{aligned}$$

APPENDIX B

SECOND APPENDIX

B.1 Posterior Inference

We develop a fully Bayesian approach for the Single-Index Model by

$$Y = \mathcal{B}_1(X^{\mathrm{T}}\boldsymbol{\theta})\boldsymbol{\beta} + \mathcal{B}_2(X^{\mathrm{T}}\boldsymbol{\gamma})\exp(\boldsymbol{\xi})\epsilon. \tag{B.1}$$

We have already specified the prior for all the parameters $(\boldsymbol{\theta}, \boldsymbol{\gamma})$, P-spline coefficients $(\boldsymbol{\beta}, \boldsymbol{\xi})$ and smoothing parameters (ρ_1, ρ_2) in Section 3.1. The prior for the error distribution parameters are specified in Section 3.2. Recall, $\widetilde{\mathbf{X}}_{\boldsymbol{\theta}} = \mathcal{B}_1(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})$, $\widetilde{\mathbf{X}}_{\boldsymbol{\gamma}} = \mathcal{B}_2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})$. We use the notation $[A]$ and $[A|B]$ to represent the marginal and conditional densities respectively. We denote $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}, X, \boldsymbol{\gamma}, \boldsymbol{\xi}, \rho_1, \rho_2) = [Y|X, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \rho_1, \rho_2]$ as the likelihood. Then the joint posterior density is

$$[\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \rho_1, \rho_2|Y, X] \propto [Y|X, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \rho_1, \rho_2][\boldsymbol{\beta}|\boldsymbol{\theta}, \rho_1][\rho_1][\boldsymbol{\theta}][\boldsymbol{\gamma}][\boldsymbol{\xi}|\boldsymbol{\gamma}, \rho_2][\rho_2].$$

B.1 Normal Error Distribution

Consistent estimation for both the mean function and the variance function can be easily attained if we assume the distribution of the error to be standard normal. So in the first section of Bayesian estimation, we consider standard normal distribution for estimating both the regression function and the variance function. We also define $\mathcal{V}_{\boldsymbol{\gamma},\boldsymbol{\xi}} = \mathrm{diag}\{\widetilde{\mathbf{X}}_{\boldsymbol{\gamma}}\exp(\boldsymbol{\xi})\}$ for notation simplicity.

The complete conditionals for $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\xi}$ require Metropolis-Hastings step. Fixed positive tuning parameters for Metropolis Hastings are denoted as δ_1, δ_2 and δ_3. The

starting values for $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, $\boldsymbol{\xi}$ are set at $\widehat{\boldsymbol{\theta}}_{\text{PHD}}, \widehat{\boldsymbol{\gamma}}_{\text{PHD}}, \widehat{\boldsymbol{\beta}}_{\text{LS}}$ and $\widehat{\boldsymbol{\xi}}_{\text{LS}}$. The values of ρ_1 and ρ_2 are initiated at a large value say, 10 to generate observations from each of the conditional distributions.

1. A candidate observation of $\boldsymbol{\theta}_{-1}$, denoted by $\boldsymbol{\theta}_{-1new}$ is generated from a proposal multivariate normal distribution centered at $\boldsymbol{\theta}_{-1current}$ and covariance matrix $\delta_1 \boldsymbol{I}_{p-1}$. With a symmetric proposal distribution, we accept $\boldsymbol{\theta}_{-1new}$ with probability $\min(1, \Delta_1)$ where,

$$\begin{aligned}
\Delta_1 &= \frac{\mathcal{L}(\boldsymbol{\theta_{new}}, \boldsymbol{\beta}, X, \boldsymbol{\gamma}, \boldsymbol{\xi}, \rho_1, \rho_2) \times p_0(\boldsymbol{\theta_{new}})}{\mathcal{L}(\boldsymbol{\theta_{current}}, \boldsymbol{\beta}, X, \boldsymbol{\gamma}, \boldsymbol{\xi}, \rho_1, \rho_2) \times p_0(\boldsymbol{\theta_{current}})} \\
&= \frac{\exp\{-(Y - \widetilde{\mathbf{X}}_{\boldsymbol{\theta_{new}}}\boldsymbol{\beta})^{\mathrm{T}} \mathcal{V}_{\boldsymbol{\gamma},\boldsymbol{\xi}}^{-1} (Y - \widetilde{\mathbf{X}}_{\boldsymbol{\theta_{new}}}\boldsymbol{\beta})/2\}}{\exp\{-(Y - \widetilde{\mathbf{X}}_{\boldsymbol{\theta_{current}}}\boldsymbol{\beta})^{\mathrm{T}} \mathcal{V}_{\boldsymbol{\gamma},\boldsymbol{\xi}}^{-1} (Y - \widetilde{\mathbf{X}}_{\boldsymbol{\theta_{current}}}\boldsymbol{\beta})/2\}} \\
&\quad \times \frac{\exp\{-(\boldsymbol{\theta_{new}} - \boldsymbol{\theta}_{\text{prior}})^{\mathrm{T}} \Sigma_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta_{new}} - \boldsymbol{\theta}_{\text{prior}})/2\}}{\exp\{-(\boldsymbol{\theta_{current}} - \boldsymbol{\theta}_{\text{prior}})^{\mathrm{T}} \Sigma_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta_{current}} - \boldsymbol{\theta}_{\text{prior}})/2\}}.
\end{aligned}$$

2. Define $A_1(\boldsymbol{\theta}, \rho_1, \boldsymbol{\gamma}, \boldsymbol{\xi}) = \left(\widetilde{X}_{\boldsymbol{\theta}}^{\mathrm{T}} \mathcal{V}_{\boldsymbol{\gamma},\boldsymbol{\xi}}^{-1} \widetilde{X}_{\boldsymbol{\theta}} + \rho_1 D\right)^{-1}$. The complete conditional distribution for the P-spline coefficient for the mean function $\boldsymbol{\beta}$ and the smoothing parameter ρ_1 are

$$\begin{aligned}
&\boldsymbol{\beta}|Y, X, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \rho_1 \sim \text{Normal}\big\{A_1(\boldsymbol{\theta}, \rho_1, \boldsymbol{\gamma}, \boldsymbol{\xi}) \widetilde{X}_{\boldsymbol{\theta}}^{\mathrm{T}} \mathcal{V}_{\boldsymbol{\gamma},\boldsymbol{\xi}}^{-1} Y, A_1(\boldsymbol{\theta}, \rho_1, \boldsymbol{\gamma}, \boldsymbol{\xi})\big\}, \\
&\rho_1|\boldsymbol{\beta} \sim \text{G}\Big\{A_{\rho_1} + M/2, \left(1/B_{\rho_1} + \boldsymbol{\beta}^{\mathrm{T}} D \boldsymbol{\beta}/2\right)^{-1}\Big\}.
\end{aligned}$$

   If we take an s^{th} order B-spline with K number of knots then M is equal to $K + s - 2$.

3. Recall that $\mathcal{V}_{\boldsymbol{\gamma},\boldsymbol{\xi}} = \text{diag}\{\widetilde{\mathbf{X}}_{\boldsymbol{\gamma}} \exp(\boldsymbol{\xi})\}$. A candidate for $\boldsymbol{\gamma}_{-1}$ is also obtained from a symmetric proposal multivariate normal distribution centered at $\boldsymbol{\gamma}_{-1current}$ and covariance matrix $\delta_2 \boldsymbol{I}_{p-1}$. We accept the candidate $\boldsymbol{\gamma}_{-1new}$with probability

equals to $\min(1, \Delta_2)$, where

$$\begin{aligned}
\Delta_2 &= \frac{\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}, X, \boldsymbol{\gamma_{new}}, \boldsymbol{\xi}, \rho_1, \rho_2) \times p_0(\boldsymbol{\gamma_{new}})}{\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}, X, \boldsymbol{\gamma_{current}}, \boldsymbol{\xi}, \rho_1, \rho_2) \times p_0(\boldsymbol{\gamma_{current}})} \\
&= \frac{\exp\{-(Y - \widetilde{\mathbf{X}}_{\boldsymbol{\theta}}\boldsymbol{\beta})^{\mathrm{T}} \mathcal{V}^{-1}_{\boldsymbol{\gamma_{new}}, \boldsymbol{\xi}} (Y - \widetilde{\mathbf{X}}_{\boldsymbol{\theta}}\boldsymbol{\beta})/2\}}{\exp\{-(Y - \widetilde{\mathbf{X}}_{\boldsymbol{\theta}}\boldsymbol{\beta})^{\mathrm{T}} \mathcal{V}^{-1}_{\boldsymbol{\gamma_{current}}, \boldsymbol{\xi}} (Y - \widetilde{\mathbf{X}}_{\boldsymbol{\theta}}\boldsymbol{\beta})/2\}} \\
&\quad \times \frac{\exp\{-(\boldsymbol{\gamma_{new}} - \boldsymbol{\gamma}_{\text{prior}})^{\mathrm{T}} \Sigma^{-1}_{\boldsymbol{\gamma_{new}}} (\boldsymbol{\gamma_{new}} - \boldsymbol{\gamma}_{\text{prior}})/2\}}{\exp\{-(\boldsymbol{\gamma_{current}} - \boldsymbol{\gamma}_{\text{prior}})^{\mathrm{T}} \Sigma^{-1}_{\boldsymbol{\gamma_{current}}} (\boldsymbol{\gamma_{current}} - \boldsymbol{\gamma}_{\text{prior}})/2\}}.
\end{aligned}$$

4. The complete conditionals for $\boldsymbol{\xi}$ also requires Metropolis algorithm, in which we generate a candidate observation of $\boldsymbol{\xi}_{new}$ from multivariate normal prior centered at $\boldsymbol{\xi}_{current}$ and covariance matrix $\delta_3 I_M$. The probability of accepting the new candidate for $\boldsymbol{\xi}$ is $\min(1, \Delta_3)$

$$\begin{aligned}
\Delta_3 &= \frac{\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}, X, \boldsymbol{\gamma}, \boldsymbol{\xi_{new}}, \rho_1, \rho_2) \times p_0(\boldsymbol{\xi_{new}})}{\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}, X, \boldsymbol{\gamma}, \boldsymbol{\xi_{current}}, \rho_1, \rho_2) \times p_0(\boldsymbol{\xi_{current}})} \\
&= \frac{\exp\{-(Y - \widetilde{\mathbf{X}}_{\boldsymbol{\theta}}\boldsymbol{\beta})^{\mathrm{T}} \mathcal{V}^{-1}_{\boldsymbol{\gamma}, \boldsymbol{\xi_{new}}} (Y - \widetilde{\mathbf{X}}_{\boldsymbol{\theta}}\boldsymbol{\beta})/2\}}{\exp\{-(Y - \widetilde{\mathbf{X}}_{\boldsymbol{\theta}}\boldsymbol{\beta})^{\mathrm{T}} \mathcal{V}^{-1}_{\boldsymbol{\gamma}, \boldsymbol{\xi_{current}}} (Y - \widetilde{\mathbf{X}}_{\boldsymbol{\theta}}\boldsymbol{\beta})/2\}} \\
&\quad \times \frac{\rho_2 \exp(\boldsymbol{\xi_{new}})^{\mathrm{T}} D \exp(\boldsymbol{\xi_{new}})}{\rho_2 \exp(\boldsymbol{\xi_{current}})^{\mathrm{T}} D \exp(\boldsymbol{\xi_{current}})}.
\end{aligned}$$

And the complete conditional for ρ_2 is

$$\rho_2 | \boldsymbol{\xi} \sim \mathrm{G}\Big[A_{\rho_2} + M/2, \big\{1/B_{\rho_2} + \exp(\boldsymbol{\xi})^{\mathrm{T}} D \exp(\boldsymbol{\xi})/2\big\}^{-1}\Big].$$

Since the value of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are continually changing in each iteration, $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}$ and $\mathbf{X}^{\mathrm{T}}\boldsymbol{\gamma}$ must be recomputed for each iteration of the MCMC in order to update the P-spline coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$. We keep track of the value of $\widehat{m} = \widetilde{\mathbf{X}}_{\boldsymbol{\theta}}\boldsymbol{\beta}$ and $\widehat{s} = \widetilde{\mathbf{X}}_{\boldsymbol{\gamma}} \exp(\boldsymbol{\xi})$ for a fixed uniformly distributed grid of points G. This enables us to keep track of

pointwise moments and percentiles. The pointwise mean curve is a natural estimate of the regression mean function $m(\cdot)$. Similarly, variance function can be estimated by pointwise variance curve.

B.2 Dirichlet Process of Infinite Mixture of Normals

We used the Blocked Gibbs Sampler (Ishwaran and James, 2001) to draw the posterior inference for the parameters specific to DPMMs and other parameters are updated using the Metropolis Hastings algortihm. For a given data set, the sample size $= n$ can acts as an upper bound on the number of mixture components in the sample for fitting DPMM. The generic notation $q(current \rightarrow proposed)$ will denote the proposal distributions of the Metropolis-Hastings steps proposing a move from the *current* value to the *proposed* value. In this paper we used a finite number of labels w. Define cluster labels as $\mathbf{C_{1:w}}$ where $C_i = k$ if ϵ_i comes from the k^{th} Cluster of the DPMM. Define the latent variables $\mathbf{Z_{1:w}}$ corresponds to the cluster level, where $Z_i = k$ if ϵ_i comes from C_k. We also define, $u_i(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}) = (y_i - \widetilde{\mathbf{X}}_{\boldsymbol{\theta}}\boldsymbol{\beta})/\sqrt{\widetilde{\mathbf{X}}_{\boldsymbol{\gamma}} \exp(\boldsymbol{\xi})}$

1. **Update the Latent Variable $Z_{1:w}$:** We propose a new value of Z_i for $i = 1, ..., w$ according to the conditional mulitnomial sampling with

$$\text{pr}(Z_i = k|\text{–}) \;=\; \frac{\pi_k\{p_k\text{Normal}(u_i; \mu_{1k}, \sigma_{1k}) + (1-p_k)\text{Normal}(u_i; \mu_{2k}, \sigma_{2k})\}}{\sum_{l=1}^{w} \pi_l\{p_l\text{Normal}(u_i; \mu_{1l}, \sigma_{1l}) + (1-p_l)\text{Normal}(u_i; \mu_{2l}, \sigma_{2l})\}}.$$

2. **Update the Stick breaking weight π_k:** We draw π_k from the marginalized conditional distribution of

$$\text{pr}(\pi_k|\text{–}) = \text{Beta}(1 + n_k, \alpha + \sum_{l=k+1}^{w} n_l), \quad k = 1, ..., w-1$$

3. **Update the constraint parameters of the DPMM** $(p, \mu, \sigma_1, \sigma_2)$**:** For all k in $Z_{i=1}^{w}$, we propose a new value for $(p_k, \mu_k, \sigma_{1k}, \sigma_{2k})$ with proposal $q\{(p_k, \mu_k, \sigma_{1k}, \sigma_{2k}) \to (p_{k,new}, \mu_{k,new}, \sigma_{1k,new}, \sigma_{2k,new})\} = \text{TU}(p_{k,new}|p_k, [0,1])$ $\text{Unif}(\sigma_{1k,new}|\sigma_{1k}, \delta_1)$ $\text{Unif}(\sigma_{2k,new}|\sigma_{2k}, \delta_2)$ $\text{Normal}(\mu_{k,new}|\mu_k, \delta_\mu)$, where $\text{TU}(\cdot|c, [a,b])$ denotes Truncated Uniform centered at c and output restricted to the interval $[a,b]$. We update the proposed value with the probability

$$\min\Big[1, \quad \frac{q\{(p_k, \mu_k, \sigma_{1k}, \sigma_{2k}) \to (p_{k,new}, \mu_{k,new}, \sigma_{1k,new}, \sigma_{2k,new})\}}{q\{(p_{k,new}, \mu_{k,new}, \sigma_{1k,new}, \sigma_{2k,new}) \to (p_k, \mu_k, \sigma_{1k}, \sigma_{2k})\}} \times \frac{\prod_{i=1}^{n} f_\epsilon\{u_i|(p_{k,new}, \mu_{k,new}, \sigma_{1k,new}, \sigma_{2k,new})\}}{\prod_{i=1}^{n} f_\epsilon\{u_i|(p_k, \mu_k, \sigma_{1k}, \sigma_{2k})\}}\Big].$$

4. **Update the Single Index parameter** $\boldsymbol{\theta}$ **for mean function** m**:** We use Metropolis Hasting Sampler to update $\boldsymbol{\theta}$ with random walk proposal for $q(\boldsymbol{\theta}_{2:p} \to \boldsymbol{\theta}_{2:p,new}) = \text{MVN}(\boldsymbol{\theta}_{2:p,new}|\boldsymbol{\theta}_{2:p}, W_{\boldsymbol{\theta}})$. We denote $\boldsymbol{\theta}_{new} = (1, \theta_{2:p,new})$ and update $\boldsymbol{\theta}$ to the proposed value with probability

$$\min\Big[1, \quad \frac{q(\boldsymbol{\theta}_{new} \to \boldsymbol{\theta})}{q(\boldsymbol{\theta} \to \boldsymbol{\theta}_{new})} \frac{\prod_{i=1}^{n} f_\epsilon\{u_i(\boldsymbol{\theta}_{new}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi})|(p_k, \mu_k, \sigma_{1k}, \sigma_{2k})\}}{\prod_{i=1}^{n} f_\epsilon\{u_i(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi})|(p_k, \mu_k, \sigma_{1k}, \sigma_{2k})\}} \frac{p_0(\boldsymbol{\theta}_{new})}{p_0(\boldsymbol{\theta})}\Big].$$

5. **Update the P-spline coefficient** $\boldsymbol{\beta}$ **for the mean function** m**:** The full conditional of $\boldsymbol{\beta}$ is given by

$$p(\beta|-) \propto p_0(\boldsymbol{\beta}|\rho_1) \prod_{i=1}^{n} f_\epsilon\{u_i(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi})|(p_k, \mu_k, \sigma_{1k}, \sigma_{2k})\}.$$

We use Metropolis-Hastings sampler to update β with random walk proposal $q(\boldsymbol{\beta} \to \boldsymbol{\beta}_{new}) = \text{MVN}(\boldsymbol{\beta}_{new}|\boldsymbol{\beta}, W_{\boldsymbol{\beta}})$. We update the smoothing hyper-parameter

ρ_1 using its closed form full conditional

$$\rho_1|\boldsymbol{\beta} \sim \mathrm{G}\Big\{A_{\rho_1} + M/2, \big(1/B_{\rho_1} + \boldsymbol{\beta}^{\mathrm{T}} D\boldsymbol{\beta}/2\big)^{-1}\Big\}.$$

6. **Update the variance function:** We define a random walk proposal $q(\boldsymbol{\gamma_{2:p}} \to \boldsymbol{\gamma_{2:p,new}}) = \mathrm{MVN}(\boldsymbol{\gamma_{new}}|\boldsymbol{\gamma}, W_{\boldsymbol{\gamma}})$ and define $\boldsymbol{\gamma_{new}} = (1, \boldsymbol{\gamma_{2:p,new}})$. Then we update $\boldsymbol{\gamma}$ to the proposed one with probability

$$\min\Big[1, \quad \frac{q(\boldsymbol{\gamma}_{new} \to \boldsymbol{\gamma})}{q(\boldsymbol{\gamma} \to \boldsymbol{\gamma}_{new})} \frac{\prod_{i=1}^n f_\epsilon\{u_i(\boldsymbol{\gamma}_{new}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\xi})|(p_k, \mu_k, \sigma_{1k}, \sigma_{2k})\}}{\prod_{i=1}^n f_\epsilon\{u_i(\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\xi})|(p_k, \mu_k, \sigma_{1k}, \sigma_{2k})\}} \frac{p_0(\boldsymbol{\gamma}_{new})}{p_0(\boldsymbol{\gamma})}\Big].$$

The full conditional of $\boldsymbol{\xi}$ is given by

$$p(\boldsymbol{\xi}|-) \propto p_0(\boldsymbol{\xi}|\rho_2) \prod_{i=1}^n f_\epsilon\{u_i(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi})|(p_k, \mu_k, \sigma_{1k}, \sigma_{2k})\}.$$

We use Metropolis-Hastings sampler to update $\boldsymbol{\beta}$ with random walk proposal $q(\boldsymbol{\xi} \to \boldsymbol{\xi}_{new}) = \mathrm{MVN}(\boldsymbol{\xi}_{new}|\boldsymbol{\xi}, W_\xi)$. Finally, we update the smoothing hyperparameter ρ_2 using its closed form full conditional

$$\rho_2|\boldsymbol{\xi} \sim \mathrm{G}\Big[A_{\rho_2} + M/2, \big\{1/B_{\rho_2} + \exp(\boldsymbol{\xi})^{\mathrm{T}} D \exp(\boldsymbol{\xi})/2\big\}^{-1}\Big].$$

The covariance matrix $W_{\boldsymbol{\theta}}, W_{\boldsymbol{\gamma}}$ of the proposal distribution for $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ is taken to be the indentity matrix multiplied by a tuning parameter. The tuning parameter is such chosen so that we can get good acceptance rates for the Metropolis- Hastings samplers, the values $\delta = 0.01, 0.1, 1$ works well for the simulation considered.

APPENDIX C

THIRD APPENDIX

C.1 Exact Likelihood: Under Normal Case for Rare Disease

Write $\Theta = (\kappa, \boldsymbol{\alpha}_x, \alpha_y, \boldsymbol{\theta}, \boldsymbol{\gamma})$. If we assume that Y given $\boldsymbol{X}$ is Normal$\{m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}), s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\}$, then the proposed likelihood function is

$$\frac{\exp\{d(\kappa + Y\alpha_y + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\alpha}_{\boldsymbol{X}})\}\phi\{(Y - m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})/s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\}}{\int\{1 + \exp(\kappa + t\alpha_y + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\alpha}_{\boldsymbol{X}})\}\phi\{(t - m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})/s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\}dt}$$

Define

$$C(\boldsymbol{X}, \Theta) \;=\; \exp\{\kappa + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\alpha}_{\boldsymbol{X}} + \alpha_y m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}) + \alpha_y^2 s^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})/2\};$$

Then the denominator of the likelihood function is

$$\begin{aligned}
& \int \exp(1 + \kappa + t\alpha_y + \boldsymbol{X}^{\mathrm{T}}\alpha_{\boldsymbol{X}})\phi\{(t - m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})/s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\}dt \\
= \;& \int s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\phi(z)dz + \int \exp\{\kappa + \boldsymbol{X}^{\mathrm{T}}\alpha_{\boldsymbol{X}} + \alpha_y m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}) + z\alpha_y s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\}s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\phi(z)dz \\
= \;& s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})[1 + \exp\{\kappa + \boldsymbol{X}^{\mathrm{T}}\alpha_{\boldsymbol{X}} + \alpha_y m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})\} \int \exp\{z\alpha_y s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\}\phi(z)dz] \\
= \;& s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})[1 + \exp\{\kappa + \boldsymbol{X}^{\mathrm{T}}\alpha_{\boldsymbol{X}} + \alpha_y m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}) + \alpha_y^2 s^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})/2\}] \\
= \;& s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\{1 + C(\boldsymbol{X}, \Theta)\}
\end{aligned}$$

Hence the loglikelihood function is

$$\begin{aligned}
\mathcal{L}(D, Y, \boldsymbol{X}, \Theta) \;=\; & -\log\{1 + c(\boldsymbol{X}, \Theta)\} - \log\{s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\} + D(\kappa + \alpha_y Y + \boldsymbol{X}^{\mathrm{T}}\alpha_{\boldsymbol{X}}) + \\
& \log[\phi\{Y - m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})/s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\}]
\end{aligned}$$

The sum of the loglikelihood function is

$$\begin{aligned}\sum_{i=1}^{n}\mathcal{L}(d_i,y_i,\boldsymbol{X}_i,\Theta) &= -\sum_{i=1}^{n}\log\{1+c(\boldsymbol{X}_i,\Theta)\}-\sum_{i=1}^{n}\log\{s({\boldsymbol{X}_i}^{\mathrm{T}}\boldsymbol{\gamma})\}\\ &\quad+\sum_{i=1}^{n}d_i(\kappa+\alpha_y y_i+{\boldsymbol{X}_i}^{\mathrm{T}}\alpha_{\boldsymbol{X}})+\\ &\quad\sum_{i=1}^{n}\log[\phi\{y_i-m({\boldsymbol{X}_i}^{\mathrm{T}}\boldsymbol{\theta})/s({\boldsymbol{X}_i}^{\mathrm{T}}\boldsymbol{\gamma})\}]\end{aligned}$$

C.2 Exact Likelihood: Under Finite Mixture of Normals for Rare Disease

Define $\boldsymbol{\Psi}=(\kappa,\boldsymbol{\alpha}_x,\alpha_y,\boldsymbol{\theta},\boldsymbol{\gamma},\{\mu_{k1}\}_{k=1}^{C},\{\sigma_{k1}\}_{k=1}^{C},\{\sigma_{k2}\}_{k=1}^{C})$. If we assume that Y given $\boldsymbol{X}$ is

$$\begin{aligned}f_\epsilon\{\frac{Y-m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})}{s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})}\} &= \sum_{k=1}^{C}\pi_k[p_k N\{\frac{Y-m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})}{s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})},\mu_{k1},\sigma_{k1}\}\\ &\quad+(1-p_k)N\{\frac{Y-m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})}{s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})},\mu_{k2},\sigma_{k2}\}]\end{aligned}$$

where, $p_k\mu_{k1}+(1-p_k)\mu_{k2}=0$ for $k=1,\cdots,C$. Then the proposed likehood function is

$$\frac{\exp\{d(\kappa+Y\alpha_y+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\alpha}_{\boldsymbol{X}})\}f_\epsilon\{(Y-m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})/s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\}}{\int\{1+\exp(\kappa+t\alpha_y+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\alpha}_{\boldsymbol{X}})\}f_\epsilon\{(t-m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})/s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\}dt}$$

Define

$$\begin{aligned}A(\boldsymbol{X},\Psi) &= \sum_{k=1}^{C}\pi_k\{p_k\sigma_{k1}+(1-p_k)\sigma_{k2}\}\\ B(\boldsymbol{X},\Psi) &= \exp\{\kappa+\boldsymbol{\alpha}_{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{X}+\alpha_Y m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})\}\\ D(\boldsymbol{X},\Psi) &= \sum_{k=1}^{C}\pi_k[p_k\sigma_{k1}\exp\{\alpha_Y s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\mu_{k1}+s^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\sigma_{k1}^2/2\}+\\ &\qquad(1-p_k)\sigma_{k2}\exp\{\alpha_Y s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\mu_{k2}+s^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\sigma_{k2}^2/2\}]\end{aligned}$$

Then the denominator of the likelihood function is

$$s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\{A(\boldsymbol{X},\Psi)+B(\boldsymbol{X},\Psi)D(\boldsymbol{X},\Psi)\}$$

Hence the loglikelihood function is

$$\begin{aligned}\mathcal{L}(D,Y,\boldsymbol{X},\Theta) &= -\log\{A(\boldsymbol{X},\Psi)+B(\boldsymbol{X},\Psi)D(\boldsymbol{X},\Psi)\}-\log\{s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\}\\ &\quad +D(\kappa+\alpha_y Y+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\alpha}_{\boldsymbol{X}})+\log[f_\epsilon\{Y-m(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta})/s(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\gamma})\}]\end{aligned}$$